Received: 26 July 2020 | Revised: 16 March 2021 | Accepted: 9 April 2021

DOI: 10.1002/int.22449

RESEARCH ARTICLE

WILEY

Learning to learn by yourself: Unsupervised meta-learning with self-knowledge distillation for COVID-19 diagnosis from pneumonia cases

Wenbo Zheng[1,2] | **Lan Yan**[2,3] | **Chao Gou**[4] | **Zhi-Cheng Zhang**[5] | **Jun J. Zhang**[2,6] | **Ming Hu**[7] | **Fei-Yue Wang**[2]

[1]School of Software Engineering, Xi'an Jiaotong University, Xi'an, China

[2]The State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China

[3]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

[4]School of Intelligent Systems Engineering, Sun Yat-sen University, Guangzhou, China

[5]Seventh Medical Center, General Hospital of People's Liberation Army, Beijing, China

[6]School of Electrical Engineering and Automation, Wuhan University, Wuhan, China

[7]Intensive Care Unit, Wuhan Pulmonary Hospital, Wuhan, China

Correspondence
Fei-Yue Wang, The State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.
Email: feiyue.wang@ia.ac.cn

Abstract

The goal of diagnosing the coronavirus disease 2019 (COVID-19) from suspected pneumonia cases, that is, recognizing COVID-19 from chest X-ray or computed tomography (CT) images, is to improve diagnostic accuracy, leading to faster intervention. The most important and challenging problem here is to design an effective and robust diagnosis model. To this end, there are three challenges to overcome: (1) The lack of training samples limits the success of existing deep-learning-based methods. (2) Many public COVID-19 data sets contain only a few images without fine-grained labels. (3) Due to the explosive growth of suspected cases, it is *urgent* and *important* to diagnose not only COVID-19 cases but also the cases of other types of pneumonia that are similar to the symptoms of COVID-19. To address these issues, we propose a novel framework called *Unsupervised Meta-Learning with Self-Knowledge Distillation* to address the problem of differentiating COVID-19 from pneumonia cases. During training, our model cannot use any true labels and aims to gain the ability of learning to learn by itself. In particular, we first present a deep diagnosis

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 61533019, 61806198, U1811463; Key Research and Development Program of Guangzhou, Grant/Award Number: 202007050002; National Key Research and Development Program, Grant/Award Number: 2018AAA0101502; National Key R&D Program of China, Grant/Award Number: 2020YFB1600400

model based on a relation network to capture and memorize the relation among different images. Second, to enhance the performance of our model, we design a self-knowledge distillation mechanism that distills knowledge within our model itself. Our network is divided into several parts, and the knowledge in the deeper parts is squeezed into the shallow ones. The final results are derived from our model by learning to compare the features of images. Experimental results demonstrate that our approach achieves significantly higher performance than other state-of-the-art methods. Moreover, we construct a new COVID-19 pneumonia data set based on text mining, consisting of 2696 COVID-19 images (347 X-ray + 2349 CT), 10,155 images (9661 X-ray + 494 CT) about other types of pneumonia, and the fine-grained labels of all. Our data set considers not only a bacterial infection or viral infection which causes pneumonia but also a viral infection derived from the influenza virus or coronavirus.

KEYWORDS
biomedical imaging, COVID-19, knowledge distillation, meta-learning, unsupervised learning

1 | INTRODUCTION

The pandemic of the coronavirus disease 2019 (COVID-19) has brought unprecedented disaster to the life of humans. Facing the ongoing outbreak of COVID-19, viral nucleic acid diagnosis using real-time polymerase chain reaction (RT-PCR) is the accepted standard diagnostic method to find COVID-19 infected people.[1-3] However, due to political and economic reasons, many hyperendemic regions and countries cannot use the RT-PCR method to identify tens of thousands of suspected patients.[4-6] To solve the lack of reagents, researchers are studying how to diagnose COVID-19 from chest X-ray images or computed tomography (CT) scans.[7-9]

The great success of deep learning methods in pneumonia diagnosis tasks has inspired many researchers.[10,11] The deep-learning-based COVID-19 diagnosis methods are emerging one after another. Still, these methods often fail to work on many data sets because there are too few images of COVID-19 in many publicly available data sets, and the previous trained deep model was not trained on COVID-19 well.[12,13] Besides, it is unrealistic that a large number of doctors label chest X-ray or CT images, and a large number of patients share their images and medical records without privacy during the ongoing outbreak of COVID-19.

On the other hand, the initial symptoms of COVID-19 are similar to those of influenza pneumonia, and bacterial pneumonia.[1,14] The COVID-19, severe acute respiratory syndrome

(SARS), and middle east respiratory syndrome (MERS) are coronaviruses, and the chest X-ray and CT images of those infected with these viruses are similar.[15-17] The outbreak is still growing rapidly. Besides, more and more COVID-19, SARS- and MERS-related CoVs were identified in animal reservoirs, raising concerns for their zoonotic transmissions and pandemic potential in the future.[18,19] Due to the increase of suspected cases, it is an *urgent*, essential and significant challenge on how to design neural networks to distinguish these viruses from chest X-ray images and CT image (slices),[20,21] which is shown in Figure 1. However, many works only focus on whether the model can distinguish COVID-19 or not.

In general, there are three challenges for COVID-19 diagnosis:

- The lack of training samples limits the success of deep-learning-based methods in this task, as small data sets typically exist in most medical imaging studies.
- In the meantime, many public data sets contain a few images from COVID-19 suspected patients and do not contain the fine-grained label of these images.
- Considering the explosive growth of suspected cases amid the COVID-19 pandemic, it is *urgent* and important to diagnose not only COVID-19 cases but also the cases of other types of pneumonia that are similar to the symptoms of COVID-19.

In contrast, there are a limited number of COVID-19 images with its variations for training, and doctors are very good at recognizing them. Why can they diagnose COVID-19 images quickly and accurately with very little direct supervision or none at all? Probably because physicians can use the experience from themselves to recognize them,[22,23] and the network cannot. Moreover, is not this one of the mechanisms of meta-learning?[24] We may use this mechanism to address the issue of limited deep-learning-based models with the lack of training samples. So, *why don't we use the principle of meta-learning to build a network?*

Furthermore, faced with the threat of the COVID-19, many medical scientists around lots of countries and regions have published many documents (i.e., papers or medical reports),[10] which contain a large number of chest images infected by COVID-19 or other

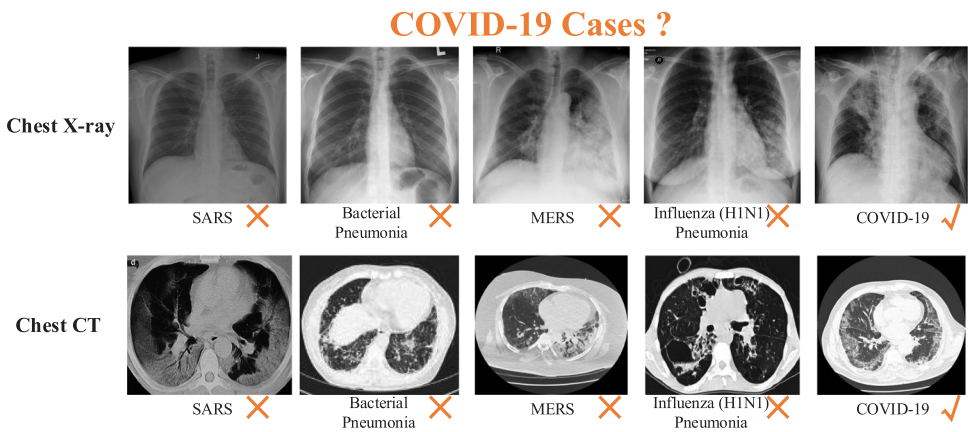


FIGURE 1 Illustration of the COVID-19 diagnosis from pneumonia cases. We constructed a data set about COVID-19 pneumonia to evaluate model performance. The model needs to distinguish five classes: COVID-19, SARS, MERS, influenza (H1N1) pneumonia, and bacterial pneumonia. MERS, middle east respiratory syndrome; SARS, severe acute respiratory syndrome [Color figure can be viewed at wileyonlinelibrary.com]

similar pneumonia. Therefore, *why does not use the images of these documents to construct a large data set?* On the other hand, considering that different countries and regions have different standards for identifying COVID-19 suspected patients[25] and describe these images in the text in corresponding documents, we can regard these text descriptions as inaccurate labels (a.k.a., noisy labels).[26,27] We may use this collection method to solve the issue of few samples from COVID-19 suspected patients. In this case, to avoid the problems of inaccurate labels, let us try to imagine there is such a model that can diagnose COVID-19 from pneumonia cases without labels or only using the label as a reference in the evaluation stage. Isn't this one of the research directions of unsupervised learning models? Thus, *can we use unsupervised learning to design our model, with the purpose of avoiding the problems of inaccurate labels?*

To address the issues mentioned above, in this paper, we propose a novel unsupervised meta-learning based model for COVID-19 diagnosis from pneumonia cases. We build a two-branch relation network via unsupervised meta-learning. First, we use the network approach to do feature extraction of training images. Then, to compare the features, we design a relation model that determines if they are from matching categories or not. Finally, to enhance the performance of our model, we design a self-knowledge distillation mechanism that distills knowledge within the model itself. The network approach is divided into several parts, and the knowledge in the deeper parts is squeezed into the shallow ones. Experimental results show that our model performs better than similar works and has strong robustness for not only diagnosing COVID-19 cases but also diagnosing the cases of other types of pneumonia that are similar to the symptoms of COVID-19.

Moreover, we propose a new chest X-ray and CT data set about COVID-19 and other types of pneumonia similar to the symptoms of COVID-19, which contains 12,851 images with the text-mined fine-grained disease labels during the ongoing outbreak of COVID-19, mined from the text radiological reports.

In summary, our main contributions are as follows:

* We propose a novel unsupervised meta-learning framework to achieve differentiate COVID-19 from pneumonia. *To the best of our knowledge, this is the first attempt to study the unsupervised meta-learning for this task.* Experimental results show that the proposed approach has strong robustness and outperforms existing similar methods.
* We design a novel self-knowledge distillation mechanism that is able to unify knowledge with different depth models utilizing a single model executable at different depths for facilitating COVID-19 diagnosis from pneumonia cases. The qualitative experiment demonstrates that this strategy is effective and achieves competitive performance. *To the best of our knowledge, this is the first attempt to study the COVID-19 diagnosis method based on this self-knowledge distillation.*
* We present a novel meta-learning-based approach to learn the discriminative features on the data sets. The qualitative discussion demonstrates that this strategy achieves competitive performance over other meta-based methods.
* A new data set about chest X-ray and CT is constructed for the task of COVID-19 diagnosis from pneumonia cases. This data set contains 2696 images (347 X-ray + 2349 CT) about COVID-19 pneumonia, 10,155 images (9661 X-ray + 494 CT) about other type of pneumonia that are similar to the symptoms of COVID-19, and the fine-grained labels of all. *To the best of our knowledge, our proposed data set is the largest data set compared to existing publicly available COVID-19 data sets except for normal cases.*

2 | RELATED WORK

We review the related work in three research streams: COVID-19 cases diagnosis and few-shot learning, knowledge distillation, self-supervised learning with pseudo labeling, and unsupervised meta-learning for image classification.

Previous COVID-19 cases diagnosis: Radiological diagnosis is a conveniently medical technique for patients who are suspected of COVID-19 in urgent need of diagnosis in serious areas. X-ray and CT scans are widely used to provide compelling evidence for the analysis of radiologists. To achieve higher accuracy for radiological diagnosis, using either X-ray or CT as the acquisition method, many works have been proposed for COVID-19 diagnosis based on chest X-ray images, the classification between COVID-19 and other non-COVID-19 subjects (including other pneumonia subjects and healthy subjects) have been explored. Zhang et al.[28] propose a ResNet based model to classify COVID-19 and non-COVID-19 X-ray images for COVID-19 diagnose. They use X-ray images from 70 COVID-19 patients and 1008 non-COVID-19 pneumonia patients, and they achieve 96.0% sensitivity and 70.7% specificity along with an area under the receiver operator curve (AUC) of 0.952. Wang et al.[29] present a deep convolutional neural networks (CNNs) based architecture called COVID-Net for COVID-19 diagnosis from X-ray images. Utilizing their own self-built COVID v2.0 data set, the COVID-Net achieves the testing accuracy of 83.5%. Also, there have been efforts made for the classification of COVID-19 from non-COVID-19 based on CT scans. Jin et al.[30] build a chest CT data set consisting of 496 COVID-19 positive cases and 1385 negative cases. They propose a two-dimensional (2D) CNN-based model for lung segmentation and a COVID-19 diagnosis model. Experimental results show that the proposed model achieves a sensitivity of 94.1%, a specificity of 95.5%, and an AUC of 0.979. *In summary, lots of studies have been proposed for X-ray-based and CT-based COVID-19 diagnosis. However, most of the recent works only consider the difference between COVID-19 and non-COVID-19 with coarse categories. Still, they consider less about different patterns between pneumonia due to diverse causes in fine-grand level.*

Few-shot learning: Few-shot learning, based on meta-learning, typically uses episodic training strategies.[31,32] In each episode, the model based on meta-learning is trained on a meta-task, which can be viewed as a classification task.[33,34] During training, the tasks were randomly selected from the training data set in the episodes. During the model evaluation, the tasks were selected from a separate test data set consisting of novel classes not included in the training data set. In summary, there are three data sets: a training set, a support set, and a testing set. The support set and testing set share the same label space, but the training set has its own label space disjointed with the support/testing set. If the support set contains K labeled examples for each of C unique classes, the target few-shot problem is called C-way$K$-shot. In C-way-K-shot few-shot learning, the model based on meta-learning is trained on the model based on meta-learning is trained on some tasks sampled from the training data set, and each task contains a support set and a query set. The task contains C unique class labels, and the support set consists of K labeled data per class. Utilizing the support set, the model learns to predict the labels in the query set. After training, the model based on meta-learning is then evaluated on new tasks sampled from the test set. Like the training tasks, each new task consists of C unique class labels with K images (in the support set). However, to assess how well the meta-learner performs on new tasks, the test data set classes do not overlap with the classes in the training set. *Following the setting of few-shot learning, we design our model to solve the problem of the diagnosis of the COVID-19 cases.*

Knowledge distillation: Knowledge distillation is one of the most popular techniques used in model compression.[35] A large number of approaches have been proposed to reinforce the efficiency of student models' learning capability. In general, teacher models and student models work in their ways, respectively, and knowledge transfer flows among different models. *In contrast, student and teacher models in our proposed self-distillation method come from the same convolutional neural networks.*

Self-supervised learning with pseudo labeling: Pseudo-labeling methods, also known as self-training, is a simple kind of self-supervised learning approach[36,37] that has been successfully applied to improve the state-of-the-art of many tasks, such as: image classification,[38,39] semantic segmentation,[40] machine translation,[41] and speech recognition.[42,43] This approach relies on two roles of networks: a teacher and the other as a student.[44] The teacher is trained on pseudo labels with unlabeled images. The student is then trained on the teacher's results with their corresponding images.[45] Thanks to the abundance of pseudo labeled data and the use of regularization methods such as data augmentation, the student learns to become better than the teacher.[39] Unlike conventional pseudo-labeling, the student and teacher in our proposed method are the same. *Therefore, we focus on our whole network, instead of one student model as usual.*

Unsupervised meta-learning for image classification: The base classes in unsupervised methods has no labels. Some scholars are able to combine with the few-shot learning methods to fulfill few-shot tasks. UFLST,[46] and CACTUs[47] use clustering to make pseudo-labels for unlabeled examples, then use the pseudo-labeled data as ordinary labeled data to construct fake few-shot tasks to complete meta-training. UMTRA[48] presents that artificial C-way 1-shot tasks are generated by randomly sampling C support examples from the training set and generating C corresponding queries with augmentation. AAL[49] and ULDA[50] generalize the UMTRA and its randomly assumption to generate/enhance randomly augmented images into classes for classification tasks. *In this paper, we also follow the random assumption and this strategy.* Further, we combine the self-knowledge distillation with unsupervised meta-learning to improve the performance of meta-learner.

3 | PROPOSED COVID-19 PNEUMONIA DATA SET

In this section, we first describe how we built our proposed *COVID-19 Pneumonia Data set* and introduce the structure of our proposed data set. Then we make a comparison with existing public available COVID-19 data sets.

3.1 | Data set creation and structure

In this paper, we propose a *COVID-19 Pneumonia Data set* by collecting medical images from radiology medical reports. Our data set mainly contains two modalities of medical imaging: X-ray and CT. In detail, X-ray images and CT images of COVID-19 are collected from radiological reports published by radiology medical centers in China, Italy, and Japan.[10] All X-ray images are posteroanterior (PA) or anteroposterior (AP) views, and salient axial slices of different CT volumes are collected for CT images. Following the work of Chest-X-ray-8,[51] we use the technology of text mining and natural language processing (NLP) to get the fine-grained labels of all images from these radiology medical reports. Our metadata attributes (fine-grained labels) are shown in our supporting information.

Our proposed data set has 12,851 2D images of different types of pneumonia, including 10,008 X-ray images and 2843 CT images (slices), consisting of five classes of pneumonia with different causes, including COVID-19, SARS, MERS, influenza, and bacterial pneumonia. Besides, the severity of COVID-19 is determined according to *Diagnosis and Treatment Protocol for COVID-19 (Trial Version 7)* issued by the National Health Commission of the People's Republic of China. To simplify the research on influenza caused pneumonia, in this paper, we only use pneumonia of influenza A virus subtype H1N1 (H1N1) infected people as an example to investigate. In other words, our data sets contain five kinds of labels: COVID-19, SARS, MERS, influenza (H1N1) pneumonia, and bacterial pneumonia. *In particular*, 347 *X-ray images and* 2346 *CT images of COVID-19 have been assembled in our data set*. We divide all images into two main categories for both X-ray and CT modalities: *Bacterial Pneumonia* and *Viral Pneumonia*. Besides, *Viral Pneumonia* contains two fine-grain classes: *Influenza (H1N1) Pneumonia* and *Coronavirus Pneumonia*. The latter *Coronavirus Pneumonia* class includes *COVID-19*, *SARS*, *MERS*. As shown in Figure 2, our data set contains two kinds of images: X-ray images and CT images (slices). CT images of patients with COVID-19 demonstrate lesions in the multiple, bilateral pulmonary. Most of the lesions chiefly appear as ground-glass opacities (GGO) in both lungs.[52,53] Some lesions have a crazy-paving appearance and consolidation.[54,55] CT images of SARS patients demonstrate that the lesions are mainly distributed in the inferior segments of both lungs.[52,56] Most patients' lesions also appear in the peripheral lung bands.[57] CT images of MERS patients show lesions mostly appear in both lungs' subpleural and basal lungs and appear as multiple GGOs and consolidation.[58] CT images of H1N1 patients demonstrate lesions usually located in the inferior lobes of both lungs. The lungs appear reticulonodular with GGO. Some patients have focal lung consolidation.[59] CT images of patients with bacterial pneumonia demonstrate that the lesions usually appear as multiple patches with GGO. Most patients have an air bronchogram.[60] The distribution characteristics of the abnormalities on X-ray images about these five types of pneumonia are similar to those of CT images (slices).[52,61-73] Although the collected 2D data (e.g., X-ray images) in our proposed data set misses lots of information (original intensity level, spacing, etc.) than original volume data, considering the usage of our proposed 2D-oriented algorithm, we have tried our best to keep the original size of the images while avoiding the problem of image distortion. To verify the usefulness of our proposed data set, experienced radiologists in our team have not only manually check all images and exclude distorted images but also retain challenging images and finally form the proposed data set. Besides, experienced radiologists in our team compare the label of the patient's medical images with the results of the patient's RT-PCR and eliminate images with errors. Basic statistics for each class of our proposed data set are shown in Table 1. Different from these existing public data sets or challenges, we focus on the analysis of different patterns between types of pneumonia with diverse causes; that is why there are no normal cases in our data set. Also, we provide other attributes such as patient sex and patient age of each image in our data set.

It is worth noting that the proposed data set is an unbalanced data set, which is exactly consistent with the long-tail problem solved by meta-learning.[74-76] Thus, a meta-learning algorithm is chosen in our paper.

3.2 | Data set comparison

Review about other public data sets: For COVID-19, which is a new type of coronavirus disease across the world, it is important to collect data for machine learning applications. In recent months, several works on COVID-19 public data sets have been proposed.[10] Cohen et al.[77]

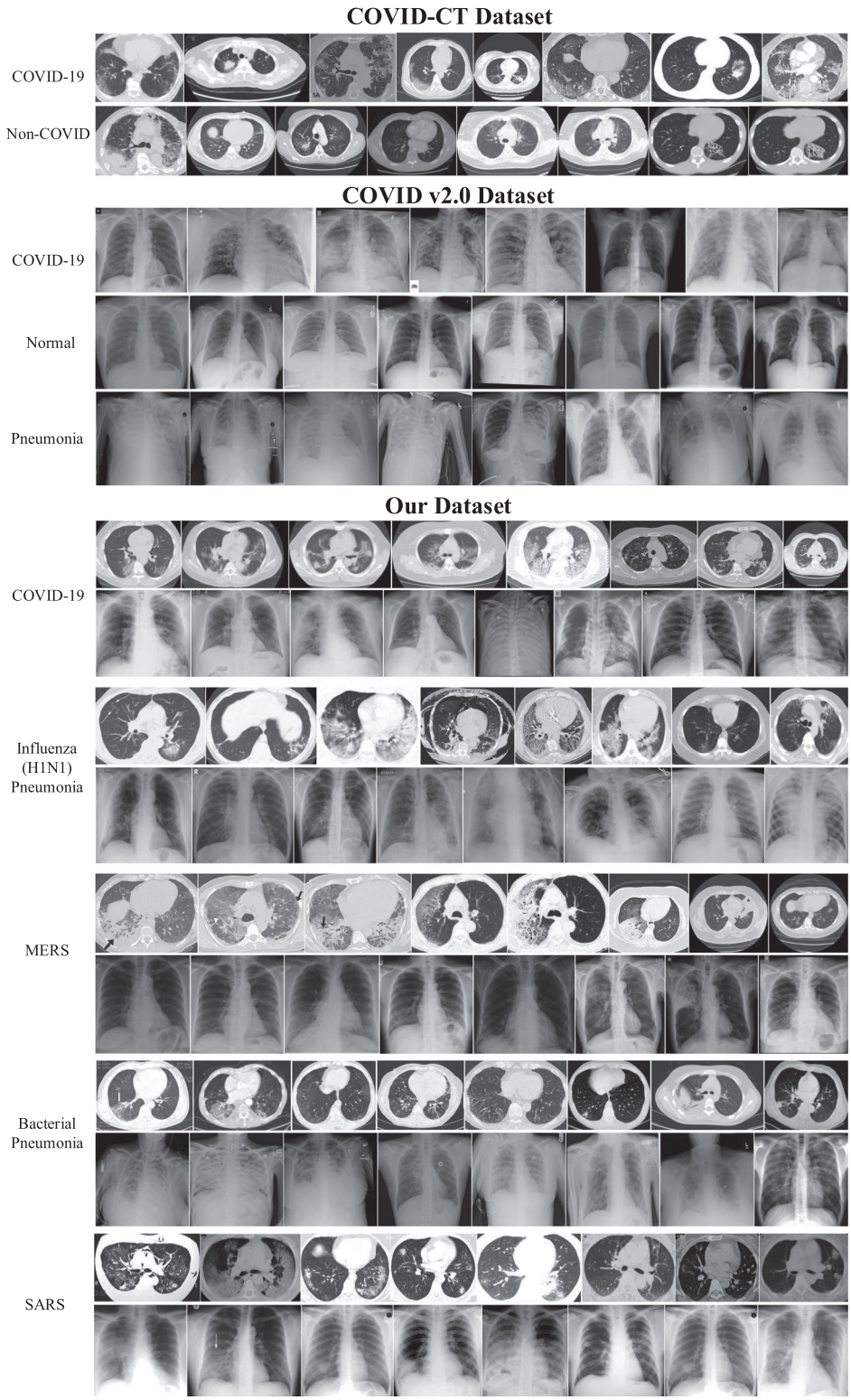


FIGURE 2 Comparable examples among three data sets

TABLE 1 Basic statistics of our proposed COVID-19 pneumonia data set

		X-ray	CT	Total
Viral pneumonia	Coronavirus pneumonia			
	COVID-19	**347**	**2349**	**2696**
	MERS	26	17	43
	SARS	48	29	77
	Influenza (H1N1) pneumonia	25	45	70
Bacterial pneumonia		9562	403	9965
	Total	10,008	2843	12,851

Abbreviations: COVID-19, coronavirus disease 2019; CT, computed tomography; MERS, middle east respiratory syndrome; SARS, severe acute respiratory syndrome.

created an image collection containing 329 images from 183 patients, most of which are chest X-ray images for COVID-19. Based on an early version of the COVID-19 image data set constructed by the above work, COVID v2.0 and its enriched version[29] added more bacterial pneumonia chest X-ray images and normal chest X-ray images. This study only contains three kinds of labels: normal, pneumonia, and COVID-19. Besides X-ray-based image data sets, CT-based image data sets are also reported recently. Zhao et al.[78] presented a publicly available COVID-CT data set consisting of COVID-19 CT axial images collected from preprinted publications from medRxiv and bioRxiv. They extracted figures and captions then judged whether a patient is positive for COVID-19 from the associated captions. Zhao et al.'s work only contain two kinds of labels: Non-COVID-19 and COVID-19. *Existing public data sets only focus on whether the images belong to a COVID-19 patient, but this kind of work ignores most of the suspected infected people who have similar symptoms to the COVID-19, such as influenza patients and bacterial pneumonia patients.*

Since our proposed data set contains both X-ray and axial CT images, to show the advancement of our proposed data set, an X-ray-based subset of our proposed data set is compared to COVID v2.0 data set,[29] and CT-based subset is compared to COVID-CT data set.[78] From Tables 2 and 3, comparable examples of X-ray images and CT images which belong to COVID v2.0 data set, COVID-CT data set and our proposed data set respectively are shown in Figure 2. It is evident that our proposed data set is better than others, the advantages of which can be summarized as follows:

TABLE 2 Comparison of COVID v2.0 data set and X-ray subset of our proposed data set

COVID v2.0[29]		**Normal**	**Pneumonia**			**COVID-19**
		8066	8614			190
Proposed	**Normal**	**Bacterial pneumonia**	**Influenza (H1N1) pneumonia**	**SARS**	**MERS**	**COVID-19**
Dataset	–	9562	25	26	48	**347**

Abbreviations: COVID-19, coronavirus disease 2019; CT, computed tomography; MERS, middle east respiratory syndrome; SARS, severe acute respiratory syndrome.

- For either X-ray images or 2D CT images, our proposed data set can be seen as the most extensive data set compared to existing publicly available COVID-19 data sets except for normal cases.
- Our proposed data set not only considers a bacterial infection or viral infection causing pneumonia but also marks the viral infection derived from influenza virus such as H1N1 or coronavirus such as COVID-19, SARS, MERS.

4 | PROPOSED METHOD

4.1 | Problem definition

Following the self-supervised learning and the unsupervised meta-learning mentioned in Section 2, we focus on the problem of COVID-19 diagnosis from pneumonia cases as unsupervised meta-learning based classification in the training process. The model is shown in Figure 3.

Learning by oneself: Due to the setting of unsupervised training, we only have an unlabeled data set $\mathcal{U} = \{(x_i)\}_{i=1}$. We sample $\mathcal{N}$ images from $\mathcal{U}$ and assume these images are from different classes. We get the data set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{\mathcal{N}}, y_i = 1, 2, 3, ..., \mathcal{N}$. We apply the data augmentation method to this data set, and we define the style of augmentation as $Aug = \{m_s\}_{s=1}^{\mathcal{M}}$. Particularly, when there is no data augmentation, we set m_s to 0. Then, we regard the style of augmentation as part of pseudo label. Finally, we can get the new data set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{\mathcal{M}\times\mathcal{N}}, y_i = \{l_z, m_s\}, l_z = 1, 2, 3, ..., \mathcal{N}$. *All in all, through the above strategy, we obtain a data set that contains pseudo labels.* In this paper, we focus on the scenario shown in Figure 1.

Learning to learn: For the task $\mathcal{T}$, the whole model contains two phases: meta-training and meta-testing. In meta-training, our training data $\mathcal{D}_{\text{meta-train}} = \{(x_i, y_i)\}_{i=1}^{\mathcal{N}_{train}}$ are used for training a classifier, where $\mathcal{N}_{train}(\mathcal{N}_{train} < \mathcal{N})$ is the number of training samples. In meta-testing, a support set of $\mathcal{N}_{support}$ labeled images $\mathcal{D}_{\text{support}} = \{(x_i, y_i)\}_{i=\mathcal{N}_{train}+1}^{\mathcal{M}\times\mathcal{N}}$. The goal is to predict the labels of a query set $\mathcal{D}_{\text{query}} = \{(x_j)\}_{j=1}^{\mathcal{N}_{query}}$, where $\mathcal{N}_{query}$ is the number of queries. Obviously, $\mathcal{D}_{\text{meta-train}}$ and $\mathcal{D}_{\text{support}}$ are from the data set $\mathcal{D}$, that is, $\mathcal{M} \times \mathcal{N} = \mathcal{N}_{train} + \mathcal{N}_{support}$. The $\mathcal{D}_{\text{query}}$ is sampled from the remaining of unlabeled data set $\mathcal{U}$. This split strategy for training and support sets is designed to simulate the support and query sets that will be encountered during test time. For traditional meta-lerning, if the support set contains K labeled examples for each of C unique classes, the target few-shot problem is called *C-way, K-shot*. In this paper, our proposed method converts the *C-way, K-shot* task to *K C-way* one-shot learning tasks. In short,

TABLE 3 Comparison of COVID-CT data set and CT subset of our proposed data set

COVID-CT[78]				COVID-19	Non-COVID-19
				349	397
Proposed	**COVID-19**	**SARS**	**MERS**	**Influenza (H1N1) pneumonia**	**Bacterialpneumonia**
Dataset	**2349**	29	17	45	403

Abbreviations: COVID-19, coronavirus disease 2019; CT, computed tomography; MERS, middle east respiratory syndrome; SARS, severe acute respiratory syndrome.

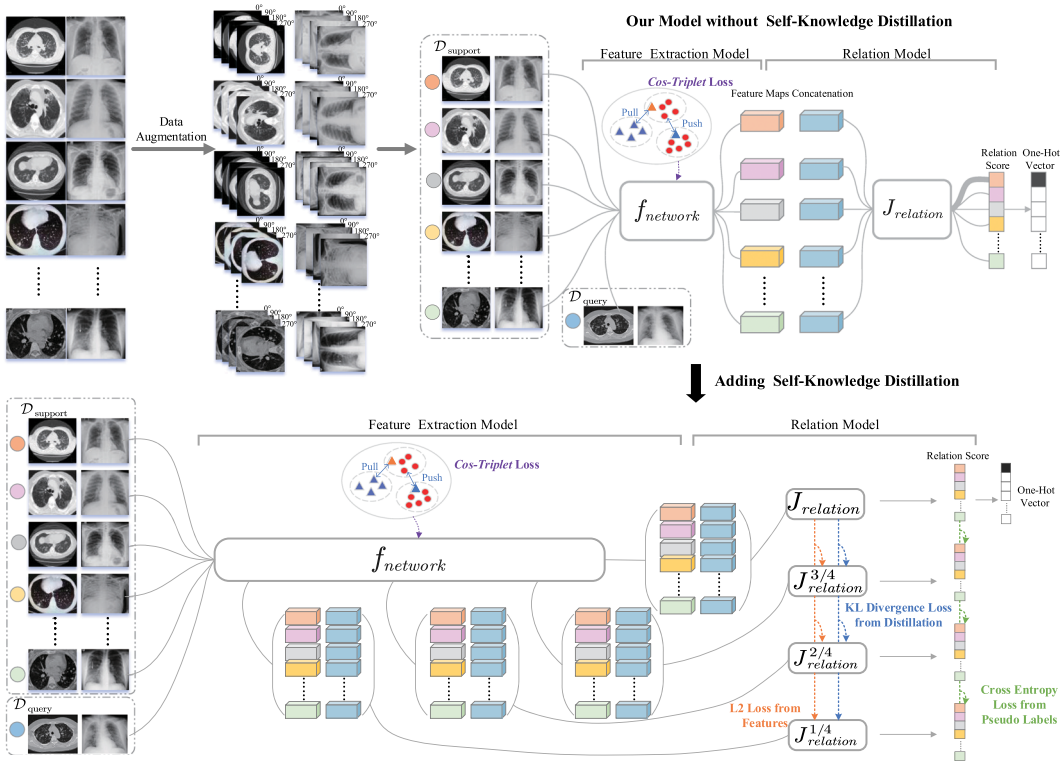


FIGURE 3 The framework of our relation network. First, we apply the data augmentation method to the data set. We use four possible 2D rotations in 0°, 90°, 180°, 270° to generate and augment this data set. Then, we build the meta-learning based model, which contains two modules: a network-based representation learning model (i.e., feature extraction model), and a relation model. The network-based representation learning model $f_{network}$ produces feature maps serving a feature extraction function. The relation model $J_{relation}(\cdot)$ represents the similarity between sample and query, which are from the training set during the training phase, and from the support set and query set, during the test phase, respectively. Further, to enhance the performance of our model, we design a self-knowledge distillation mechanism that distills knowledge within the model itself. The network-based representation learning model is divided into several parts, and the knowledge in the deeper part is squeezed into the shallow ones. In this way, we add our relation model after each part and name these as: $J^{1/4}_{relation}(\cdot)$, $J^{2/4}_{relation}(\cdot)$, $J^{3/4}_{relation}(\cdot)$ and $J_{relation}(\cdot)$. The whole model updates itself and produces the final results [Color figure can be viewed at wileyonlinelibrary.com]

we aim to learn knowledge via training our model from $\mathcal{D}_{\text{meta-train}}$, and transfer this extracted knowledge in the testing phase to classify the images in $\mathcal{D}_{\text{query}}$ given $\mathcal{D}_{\text{support}}$.

Due to the increase of new COVID-19 cases, it is a crucial issue whether the model trained based on existing data can be applied to these new cases. Therefore, we consider that (1) nonlinear mapping in artificial neural networks should be generalizable to work with samples of novel classes, and (2) the mapping should preserve the relationship between classes on the unseen class samples in $\mathcal{D}_{\text{query}}$. We propose a novel relational network to address the problem of COVID-19 diagnosis from pneumonia cases. First, we meta-learn a transferable feature extraction model based on *proposed Cos-Triplet loss* defined in Equation (1) from the training data set. The well-learned features of the query samples in the support set are then fed into the nonlinear distance metric to learn the similarity scores. Further, we conduct a few-shot

classification based on these scores. As illustrated in Figure 4, our network representation learning consists of two branches: a *feature extraction model* and a *relation model* during the training of our network.

Meta-learning based feature extraction: The training images from $\mathcal{D}_{\text{meta-train}}$ and $\mathcal{D}_{\text{support}}$ are randomly selected to form a triplet $(x_a; x_p; x_n)$ with an anchor images x_a, a positive images x_p, and a negative images x_n. The label of the selected images in a triplet should satisfy $y_a = y_p \neq y_n$. Then, we can get the anchor images set $\mathcal{X}_a$, positive images set $\mathcal{X}_p$ and negative images set $\mathcal{X}_n$ by following the above strategy. We aim to pull the feature maps of anchor and positive images close to each other, as shown in Figure 3, while pushing the feature maps of anchor and negative images far apart.

4.2 | Network representation learning

For each update in the network, the traditional triplet loss[79] only interacts with a negative image (or a negative class), and we need to compare the query images in several different classes for few-shot classification. Therefore, triplet loss may not be sufficient for feature embedding learning, especially when we need to deal with multiple classes in few-shot classification settings. Inspired by SoftTriple loss,[80] we generalize the traditional triplet loss to a novel

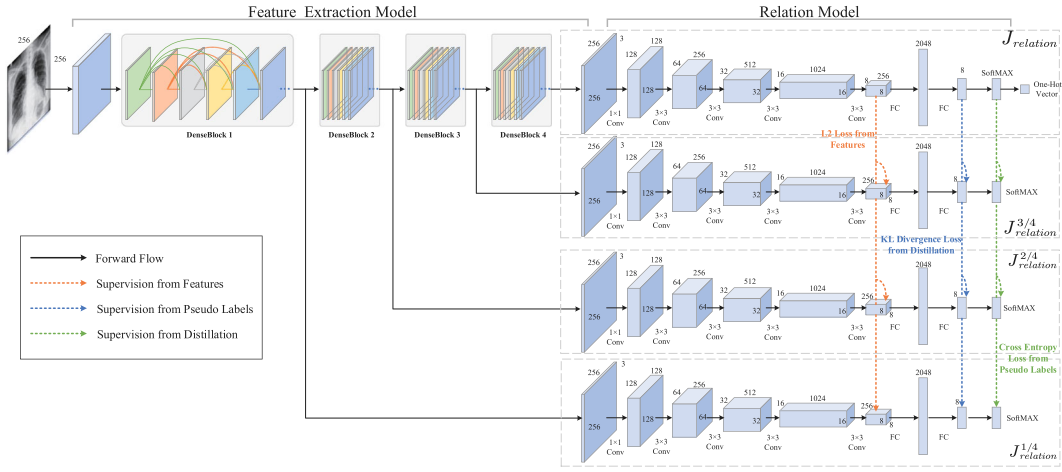


FIGURE 4 The architecture of our relation network. First of all, we use the network based on the DensetNet-121 basebone to extract the features of the given image. Second, DensetNet-121 has been divided into four parts according to its depth. We add our relation model after each part, and name these as: $J^{1/4}_{relation}(\cdot)$, $J^{2/4}_{relation}(\cdot)$, $J^{3/4}_{relation}(\cdot)$ and $J_{relation}(\cdot)$. Each part is viewed as independent with a different performance. Third, each part is trained under three kinds of supervision: (1) Supervision from the features, where we use the L2 loss from hints. In this way, inexplicit/tacit knowledge in the feature maps is brought into the bottleneck layer of each shallow classifier, leading all classifiers to conform the feature maps in their bottleneck layer to the of the deepest classifier; (2) *Supervision from the pseudo labels*, we use the cross-entropy loss to not only the deepest classifier, but also all the shallow classifiers. In this way, the knowledge hidden in the data set is introduced directly from pseudo labels to all the classifiers; (3) Supervision from the distillation, where we use the Kullback–Leibler (KL) divergence to achieve the self-supervision. In this way, the deepest network as a teacher can guide each shallow classifier as students. Finally, we apply the relation model and get the final results [Color figure can be viewed at wileyonlinelibrary.com]

triplet loss, called *Cos-Triplet Loss*. Specifically, for each anchor, the loss aims to pull data of the same class close to the anchor and to push others away in the embedding space. The loss not only considers the relations among data but also associates all data with each anchor so that the gradients concerning a data point are weighted by its relative proximity to the anchor (i.e., relative hardness) affected by the other data.

All in all, the key idea is to associate each anchor with the entire data so that the data interact with each other through the anchor during training. The function $f_{network}$ represents feature extraction function using the network to produce feature maps $f_{network}(x_a), f_{network}(x_p)$ and $f_{network}(x_n)$. We design Cos-Triplet loss during the training procedure of feature extraction:

$$\begin{aligned}&\mathcal{L}_{Cos\text{-}Triplet} = \\ &\frac{1}{|\mathcal{X}_p|} \times \sum \log\left(1 + \sum_{1}^{|\mathcal{X}_a|} e^{-\alpha(\cos(f_{network}(x_p), f_{network}(x_a)) - \beta)}\right) \\ &+ \frac{1}{|\mathcal{X}_a| + |\mathcal{X}_p| + |\mathcal{X}_n|} \\ &\times \sum \log\left(1 + \sum_{1}^{|\mathcal{X}_n|} e^{\alpha(\cos(f_{network}(x_n), f_{network}(x_a)) + \beta)}\right)\end{aligned} \tag{1}$$

where $|\cdot|$ means the number of elements in the set $\cdot$,$\cos(\cdot,\cdot)$ denotes the cosine similarity between two feature maps, $\alpha > 0$ is a scaling factor, and $\beta > 0$ is a margin.

Meta-learning based relation model: We further design a nonlinear distance relation model for learning to compare the image features in a few-shot classification.

Given image $x^{support}$ in support set $\mathcal{D}_{\text{support}}$ and image x_i in the train set $\mathcal{D}_{\text{meta-train}}$, we assume the $C_{network}(\cdot,\cdot)$ to be concatenation of the corresponding feature maps of the two images at the same depth. The combined feature map of images from the support set and train set is used as the relation model $J_{relation}(\cdot)$ to get a scalar in range of 0 to 1 representing the similarity between x_i and $x^{support}$, which is called relation score. Suppose we have one labeled sample from the train set, our model can generate $\mathcal{N}_{train}$ relation scores $Judge_i$ for the relation between one image input $x^{support}$ from support set and training image set examples x_i:

$$\begin{aligned}&Judge_i = J_{relation}(f_{network}(x_i)) = J_{relation}(C_{network}(f_{network}(x^{support}), f_{network}(x_i))) \\ &\quad i = 1, 2, \cdots, \mathcal{N}_{train}\end{aligned} \tag{2}$$

Furthermore, we can do the operation of the element-wise sum over our feature extraction model outputs of all samples from each training class to form this class's feature map. Moreover, this pooled class-level feature map is concatenated with the feature map of the test samples as above.

We use SoftMAX loss[81] to train our relation model, regressing the relation score $Judge_i$ to the pseudo label: matched pairs have similarity 1, and the mismatched pair have similarity 0.

4.3 | Data augmentation

In this paper, our data augmentation focuses on four possible 2D rotations in $Rot = \{0°, 90°, 180°, 270°\}$, denoted as $m_s \in \{0, 90, 180, 270\}$. This rotation method is a common data augmentation strategy in medical image analysis.[82,83] Specifically, given an image x_i, we

first create its four rotated copies $\mathcal{D} = \{(x_i, y_i)\}^{\mathcal{M} \times \mathcal{N}}$. Based on the features $f_{network}(x_i)$ extracted from such a rotated image, the SoftMAX classifier $J_{relation}(\cdot)$ (i.e., relation model) predict the rotation class m_s. Accordingly, the unsupervised loss of this task is defined as:

$$\mathcal{L}_{self} = \mathbb{E}\Big[\sum - \log J_{relation}(f_{network}(x_i))\Big] \tag{3}$$

4.4 | Self-knowledge distillation

We put forward a novel self-knowledge distillation as shown in Figure 4. First, the network of feature extraction model can be divided into some shallow parts according to its depth and original structure. In this paper, DenseNet-121 can be divided into four parts according to DenseBlocks. Second, a relation model, playing a role in a classifier which is only utilized in training and can be removed in testing, is set after each shallow part. In adding each relation model, the convolutional layers consider the impacts between each shallow part, and add L2 loss from its extracted features. According to knowledge distillation,[84-87] all parts with corresponding relation model can be regarded as student models, and the deepest can be regarded as the teacher model.

Relation models (the proposed self-knowledge distillation has multiple relation models within a whole network) in the neural network are denoted as $\Theta = \left\{J^{r/R}_{relation}\right\}_{r=1}^{R}$, where R (in this paper, we set $R = 4$) denotes the number of relation models. The output of each relation model $J^{r/R}_{relation}(\cdot)$ is denoted as $\hat{y}_i^{r/R}$, correspondingly. When $r = R$, we denote $\hat{y}_i^{4/R}$ as $\hat{y}_i$. Similar to knowledge distillation,[84] our model has three kinds of supervisions:

Supervision from the features: This supervision, whose goal is to guide the learning of student models, is from the features of the deepest model. It works by decreasing the distance between feature maps in each shallow model and the deepest model. It can be obtained through the computation of the L2 loss between feature maps. Using the L2 loss, the latent knowledge in feature maps is introduced to convolutional layers in each shallow part, where all feature maps in their convolutional layers fit the feature maps of the deepest model. The L2 loss is written as:

$$\mathcal{L}^r_{feature} = \mu \times ||\mathcal{F}_{r/R} - \mathcal{F}_R||_2^2 \tag{4}$$

where μ means the hyper-parameter, $\mathcal{F}_{r/R}$ and $\mathcal{F}_R$ denote features in the $J_{relation_{r/R}}$ and features in the deepest classifier $J_{relation}$, respectively.

Supervision from the distillation: The goal of the supervision is to make a shallow model approximate the deepest model. We use Kullback–Leibler (KL) divergence loss between the outputs of students and teachers. The KL loss is written as:

$$\mathcal{L}^r_{distillation} = \nu \times KL\Big(\hat{y}_i^{r/R}, \hat{y}^R\Big) \tag{5}$$

where μ means the hyper-parameter and $\hat{y}^R$ means the output of the deepest model $J_{relation}$.

Supervision from the pseudo labels: Cross entropy loss is from pseudo labels to not only all shallow models, but also the deepest model (i.e., Equation 3). It is computed with the labels from the training data set and the outputs of each model. The cross entropy loss is written as:

$$\mathcal{L}^r_{label} = (1 - \nu) \times \mathcal{L}_{CE}\Big(\hat{y}_i^{r/R}, y_i\Big) \tag{6}$$

where $\mathcal{L}_{CE}$ means a standard cross-entropy loss.[88]

Self-knowledge distillation loss: To sum up, the loss function of the whole self-knowledge distillation consists of the loss function of each supervision, which can be written as:

$$\mathcal{L}_{SKD} = \sum_{r}^{R}\left(\mathcal{L}^{r}_{feature} + \mathcal{L}^{r}_{distillation} + \mathcal{L}^{r}_{label}\right) \tag{7}$$

where R denotes the number of relation models. By experiments, we set ν and μ to $\frac{1}{2}$ and 3, respectively.

All in all, the full training procedure of self-knowledge distillation is following as:

procedure Self-Knowledge Distillation
  Initialize parameters Θ.
  while Θ has not converged **do**
    Sample a batch $\{x_i, y_i\}_{i=1}^{N_{bacth}}$ from the data set $\mathcal{D}_{\text{meta-train}}$.
    for $i = 1$ to N_{bacth} **do**
      Compute the output of each relation model with its parameters Θ:

$$\hat{y}^{r/R}{}_i = J^{r/R}_{relation}(f_{network}(x_i))$$

      Update parameters Θ by computing the gradients of the proposed loss function Equation (7)
    end for
  end while
end procedure

4.5 | Training methods

Since the proposed self-knowledge distillation distills knowledge from the current training model, at the beginning of the training process, the model does not contain relevant information. That is, we cannot extract any knowledge from the training model at the beginning. Thus, we start training process without knowledge distillation at first and gradually increase the amount of knowledge distillation as the training iteration goes. Therefore, in the first stage, our training model starts with the feature extraction function in Equation (1) and data augmentation function in Equation (3), as shown in Figure 3. After training the model for a while, we perform the second stage: it gradually transits to the loss of self-knowledge distillation in Equation (7), as shown in Figure 3. Our network architecture is depicted in Figures 3 and 4. There are two components to our network:

- Feature Extraction Model: We employ the DenseNet-121 architecture[89] for learning the feature extraction model. Note that we remove the classification layer of the original DenseNet. When meta-learn the transferable feature extraction, we use SGD with a learning rate of 0.002 and a decay for every 40 epochs. We train 800 epochs at the first stage for the losses in Equations (1) and (3), and then train 200 epochs at the second stage for the loss in Equation (7). During these processes, we adopt the semi-hard mining strategy when the loss starts to converge. We set α and β to 32 and 10^{-1}, respectively, for all experiments.
- Relation Model: We use the 8-layer network architecture. Taking a sample feature map as input, the output of the 8-th pooling layer is one-hot vector. The kernels of network change in

turns: $3 \times 256 \times 256 \rightarrow 128 \times 128 \times 128$ (Convolution, kernel size: 1×1) $\rightarrow 256 \times 64 \times 64$ (Convolution, kernel size: 3×3) $\rightarrow 512 \times 32 \times 32$ (Convolution, kernel size: 3×3) $\rightarrow 1024 \times 16 \times 16$ (Convolution, kernel size: 3×3) $\rightarrow 256 \times 8 \times 8$. Then, we apply the fully connected layer to change into 2048-dimensional vector. Finally, we use one fully connected layer and one SoftMAX layer to have 8 and 1 outputs, respectively. Other settings are similar to our feature extraction model.

5 | EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we experimentally evaluate the performance of the proposed model on two public benchmark data sets and our data set, and we compare its performance with other state-of-the-art deep representation learning models. During the evaluation of our model, we first can get predicted labels of the query set via training and testing our model. Then, we use the real label of the data set $\mathcal{D}$ to get corresponding true labels of the query set. Finally, we focus on the difference between the true meaning predicted labels and real labels of the query set.

5.1 | Experimental settings

In this section, we outline the criteria used for evaluation, and then we describe the evaluation protocol.

Evaluation criteria: We use the accuracy, precision, F1-score, sensitivity, specificity, and AUC to assess the performance of all models. More precisely, we use sensitivity and specificity to denote the proportion of positive samples and negative samples that are correctly identified, respectively. Besides, we use AUC to measure the overall classification performance, which is sensitive to the imbalance among multiple classes.

Meta-learning protocol: The classic pipeline in meta-learning is first to train a model on a set of base classes and then to evaluate it on a different set of novel classes (each set of classes is split into train and validation subsets).[90] For our experiments, we use this protocol. In this paper, we use COVID v2.0 data set,[29] COVID-CT data set[78] and our data set. In this paper, we set C as 5 for the K *C-way* one-shot learning task. Specifically, we randomly choose the 1600 images from the training images of the COVID v2.0 data set, where the number of training image is 16590, to construct the training set; we randomly choose 1510 images from the test images of the COVID v2.0 data set, where the number of the testing image is 1953, to construct the test set. We randomly choose the 6000 chest X-ray images from our data set to construct the training set; we randomly choose 2000 other chest X-ray images from our data set to construct the test set. Besides, we randomly choose the 2000 chest CT images from our data set to construct the training set; we randomly choose 500 other chest CT images from our data set to construct the test set. Note that we randomly choose 10 times as per the above strategy and take the average evaluation criteria for comparison. On the COVID-CT data set, we follow the data split of this data set. We use 425 (191 COVID-19 + 234 non-COVID-19) images to construct the training set; we use 118 (60 COVID-19 + +58 non-COVID-19) images to construct the support set; we use 203 (98 COVID-19 + 105 non-COVID-19) images to construct the query set.

5.2 | The results of our model

In this section, we show the results on the COVID v2.0 data set, COVID-CT data set, and our data set, which are in Figure 5. "Ours w/o S-KD" means a variant of Ours, which only using network representation learning and not using self-knowledge distillation.

On our data set using our chest X-ray images, the accuracy, precision, sensitivity, specificity, F1-score of ours w/o S-KD is 0.989, 0.986, 0.937, 0.998, 0.961, and 0.978. The accuracy, precision, sensitivity, specificity, F1-score of ours is 0.995, 0.997, 0.967, 0.999, 0.981, and 0.987. Ours is 0.006, 0.011, 0.030, 0.002, and 0.021 higher than ours w/o S-KD, in terms of accuracy, precision, sensitivity, specificity, and F1-score, respectively. On our data set using our chest CT images, the accuracy, precision, sensitivity, specificity, F1-score of ours w/o S-KD is 0.949, 0.984, 0.923, 0.981, and 0.952. The accuracy, precision, sensitivity, specificity, F1-Score of Ours is 0.986, 0.990, 0.987, 0.986, and 0.988. Ours is 0.037, 0.006, 0.064, 0.005, and 0.036 higher than Ours w/o S-KD, in terms of accuracy, precision, sensitivity, specificity, and F1-Score, respectively. This means our method is resultful and effective. Moreover, by analyzing others shown in Figure 5 on the COVID v2.0 data set and COVID CT data set, we can get similar conclusions.

5.3 | Comparison with state-of-the-art methods

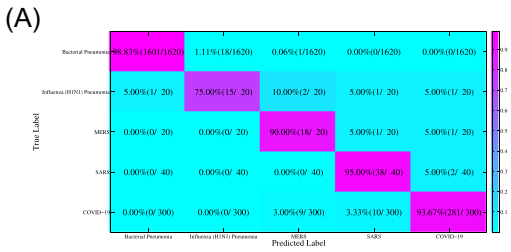
We compare the state-of-the-art approaches with our model on three data sets, including the COVID v2.0 data set, COVID-CT data set, and our data set. "Ours w/o S-KD" means a variant of ours, which only using network representation learning and not using self-knowledge distillation.

COVID-19 diagnosis from pneumonia cases on our data set: The COVID-19 diagnosis from pneumonia cases here is: given chest X-ray or CT images, we can *not only* diagnose COVID-19 *but also* identify other types of pneumonia (i.e., SARS, MERS, influenza (H1N1) pneumonia, bacterial pneumonia) that are similar to the symptoms of COVID-19.

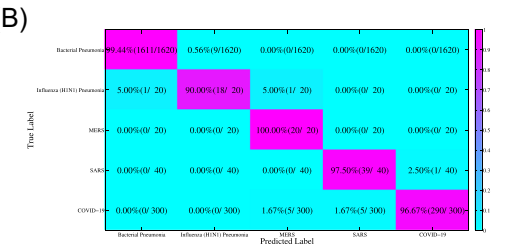
Baselines on our data set: We compare against various state-of-the-art baselines on our data set, including DenseNet-121,[91] DenseNet-161,[92] ResNet-34,[93] VGG-19,[94] ResNet-18,[95] EfficientNet-B0,[96] EfficientNet-B1,[96] EfficientNet-B2,[96] EfficientNet-B3,[96] EfficientNet-B4,[96] EfficientNet-B5,[96] Inception-v3,[97] Inception-ResNet-v2,[98] MobileNet-v2,[99] DenseNet-201,[100] and VGG-16.[94]

Effect of proposed self-knowledge distillation: For evaluating the performance of our approach, we compare results reported in row-"Ours w/o S-KD" and row-"Ours" from Table 4. Our approach leverages the same loss functions and features in row-"Ours w/o S-KD" for a fair comparison. From Table 4, we find that our approach improves performance consistently in all the cases. *It is evident that the design of self-knowledge distillation can enhance the effectiveness of our approach.*

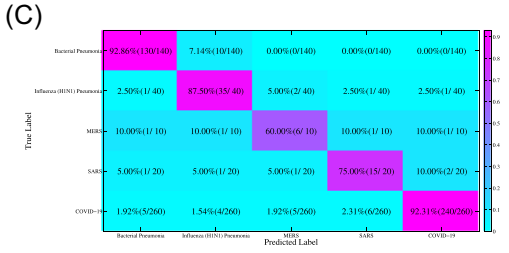
Effect of our approach: From Table 4, it is evident that our approach is better than others. Specifically, ours is 0.069, 0.181, 0.146, 0.168, 0.137, 0.194, 0.296, 0.233, 0.257, 0.233, 0.179, 0.096, 0.128, 0.131, 0.235, and 0.265 higher than DenseNet-121, DenseNet-161, ResNet-34, VGG-19, ResNet-18, EfficientNet-B0, EfficientNet-B1, EfficientNet-B2, EfficientNet-B3, EfficientNet-B4, EfficientNet-B5, Inception-v3, Inception-ResNet-v2, MobileNet-v2, DenseNet-201, VGG-16, in term of accuracy, respectively. In terms of precision, sensitivity, specificity, F1-Score, AUC, there are similar scenarios as the above. *From above, our approach is more effective and robust than the state-of-the-art approaches on our data set using chest CT images.* Besides, on our data set using chest



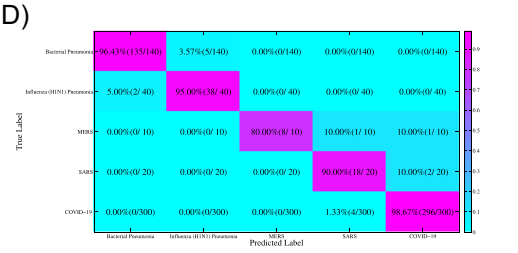
Confusion Matrix of Ours w/o S-KD on Our Chest X-ray Images



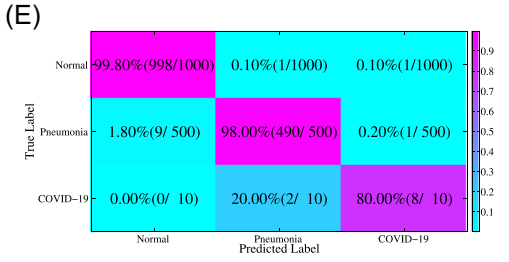
Confusion Matrix of Ours on Our Chest X-ray Images



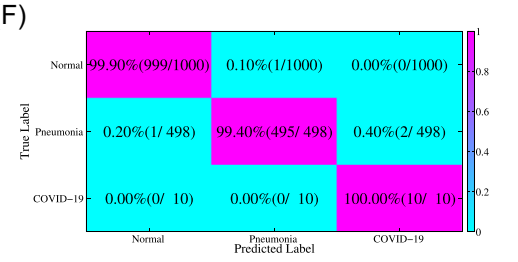
Confusion Matrix of Ours w/o S-KD on Our Chest CT Images



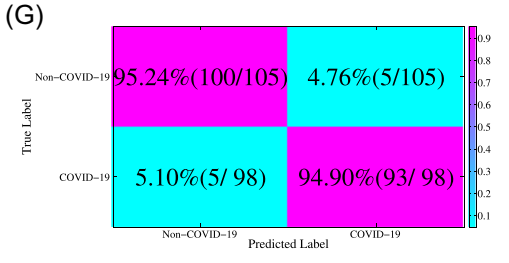
Confusion Matrix of Ours on Our Chest CT Images



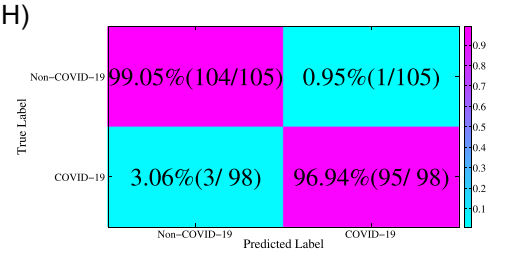
Confusion Matrix of Ours w/o S-KD on COVID v2.0 Dataset



Confusion Matrix of Ours on COVID v2.0 Dataset



Confusion Matrix of Ours w/o S-KD on COVID-CT Dataset



Confusion Matrix of Ours on COVID-CT Dataset

FIGURE 5 Confusion matrix of our model (a) Confusion matrix of ours w/o S-KD on our chest X-ray images; (b) confusion matrix of ours on our chest X-ray images; (c) Confusion matrix of ours w/o S-KD on our chest CT images; (d) Confusion matrix of ours on our chest CT images; (e) Confusion matrix of ours w/o S-KD on COVID v2.0 data set; (f) Confusion matrix of ours on COVID v2.0 data set; (g) Confusion matrix of ours w/o S-KD on COVID-CT data set; (h) Confusion matrix of ours on COVID-CT data set [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 4 Classification results of each model on our data set using chest CT images

Methods		Accuracy	Precision	Sensitivity	Specificity	F1-Score	AUC
Ours	Unsupervised	**0.986**	**0.990**	**0.987**	**0.986**	**0.988**	**0.991**
Ours w/o S-KD	Unsupervised	0.949	0.984	0.923	0.981	0.952	0.934
DenseNet-121[91]	Supervised	0.918	0.877	0.897	0.917	0.912	0.931
DenseNet-161[92]	Supervised	0.805	0.886	0.863	0.848	0.875	0.833
ResNet-34[93]	Supervised	0.840	0.848	0.865	0.851	0.856	0.829
VGG-19[94]	Supervised	0.818	0.857	0.830	0.813	0.843	0.809
ResNet-18[95]	Supervised	0.849	0.865	0.824	0.789	0.844	0.791
EfficientNet-B0[96]	Supervised	0.793	0.801	0.822	0.769	0.784	0.943
EfficientNet-B1[96]	Supervised	0.691	0.681	0.687	0.650	0.696	0.739
EfficientNet-B2[96]	Supervised	0.754	0.714	0.710	0.677	0.721	0.809
EfficientNet-B3[96]	Supervised	0.730	0.753	0.760	0.706	0.704	0.828
EfficientNet-B4[96]	Supervised	0.754	0.765	0.761	0.775	0.747	0.901
EfficientNet-B5[96]	Supervised	0.807	0.765	0.761	0.790	0.812	0.825
Inception-v3[97]	Supervised	0.890	0.838	0.850	0.875	0.890	0.922
Inception-ResNet-v2[98]	Supervised	0.858	0.877	0.817	0.813	0.891	0.910
MobileNet-v2[99]	Supervised	0.855	0.884	0.945	0.754	0.858	0.911
DenseNet-201[100]	Supervised	0.752	0.796	0.765	0.706	0.780	0.731
VGG-16[94]	Supervised	0.721	0.723	0.757	0.701	0.740	0.700

X-ray images, there are similar scenarios as the above and shown in Table 5. *These mean our approach can diagnose COVID-19 from pneumonia cases effectively and robustly.*

COVID-19 cases diagnosis on COVID v2.0 data set: The COVID-19 cases diagnosis: given chest X-ray or CT images, we can *only* diagnose COVID-19 and *do not have to* diagnose other type of pneumonia.

Baselines on COVID v2.0 data set: We compare against various state-of-the-art baselines on the COVID v2.0 data set, including COVID-CAPS1,[101] COVID-ResNet2,[102] AI4COVID-193,[103] DenseNet-121,[91] DenseNet-161,[92] ResNet-34,[93] VGG-19,[94] ResNet-18,[95] EfficientNet-B0,[96] EfficientNet-B1,[96] EfficientNet-B2,[96] EfficientNet-B3,[96] EfficientNet-B4,[96] EfficientNet-B5,[96] Inception-v3,[97] Inception-ResNet-v2,[98] MobileNet-v2,[99] DenseNet-201,[100] and VGG-16.[94]

Effect of proposed self-knowledge distillation: "Ours" is 0.001, 0.033, 0.200, 0.000, 0.109, 0.092 higher than "Ours w/o S-KD," in term of accuracy, precision, sensitivity, specificity, F1-score, AUC. These improvements once again show that learning by self-knowledge distillation for the better performance of COVID-19 case diagnosis.

Effect of our approach: From Table 6, it is visible that our approach is better than others. Specifically, ours is 0.042, 0.045, 0.110, 0.074, 0.193, 0.164, 0.181, 0.150, 0.200, 0.303, 0.240, 0.263, 0.240, 0.185, 0.101, 0.134, 0.138, 0.249, and 0.282 higher than COVID-CAPS, COVID-ResNet, AI4COVID-19, DenseNet-121, DenseNet-161, ResNet-34, VGG-19, ResNet-18, EfficientNet-B0, EfficientNet-B1, EfficientNet-B2, EfficientNet-B3, EfficientNet-B4, EfficientNet-B5, Inception-v3,

TABLE 5 Classification results of each model on our data set using chest X-ray images

Methods		Accuracy	Precision	Sensitivity	Specificity	F1-Score	AUC
Ours	Unsupervised	**0.995**	**0.997**	**0.967**	**0.999**	**0.981**	**0.987**
Ours w/o S-KD	Unsupervised	0.989	0.986	0.937	0.998	0.961	0.978
DenseNet-121[91]	Supervised	0.918	0.877	0.897	0.917	0.911	0.891
DenseNet-161[92]	Supervised	0.805	0.885	0.861	0.842	0.873	0.833
ResNet-34[93]	Supervised	0.838	0.845	0.861	0.852	0.853	0.833
VGG-19[94]	Supervised	0.819	0.852	0.837	0.818	0.844	0.816
ResNet-18[95]	Supervised	0.855	0.866	0.823	0.787	0.844	0.792
EfficientNet-B0[96]	Supervised	0.792	0.801	0.822	0.769	0.783	0.842
EfficientNet-B1[96]	Supervised	0.691	0.681	0.688	0.649	0.695	0.766
EfficientNet-B2[96]	Supervised	0.753	0.715	0.710	0.677	0.721	0.789
EfficientNet-B3[96]	Supervised	0.729	0.753	0.760	0.706	0.703	0.787
EfficientNet-B4[96]	Supervised	0.753	0.765	0.761	0.775	0.746	0.877
EfficientNet-B5[96]	Supervised	0.807	0.765	0.761	0.790	0.812	0.838
Inception-v3[97]	Supervised	0.890	0.838	0.851	0.875	0.889	0.934
Inception-ResNet-v2[98]	Supervised	0.858	0.877	0.817	0.813	0.889	0.930
MobileNet-v2[99]	Supervised	0.855	0.884	0.946	0.754	0.858	0.943
DenseNet-201[100]	Supervised	0.755	0.794	0.762	0.706	0.778	0.724
VGG-16[94]	Supervised	0.718	0.721	0.754	0.704	0.738	0.699

Inception-ResNet-v2, MobileNet-v2, DenseNet-201, and VGG-16, in term of accuracy, respectively. *From above, our approach is more effective and robust than the state-of-the-arts on the COVID v2.0 data set.* By analyzing the results shown in Table 7 on COVID CT data set, we can get similar conclusions. *These mean our approach can diagnose COVID-19 cases diagnosis effectively and robustly.*

5.4 | Compared to different self-knowledge distillation

To better verify the effective performance of our self-knowledge distillation, we compare our model with other state-of-the-art self-knowledge distillations, including CS-KD,[35] BYOT,[104] and DDGSD.[105] "Ours (CS-KD)" means a variant of Ours, which using CS-KD and not using our self-knowledge distillation. "Ours (BYOT)" means a variant of Ours, which using BYOT and not using our self-knowledge distillation. "Ours (DDGSD)" means a variant of Ours, which using DDGSD and not using our self-knowledge distillation. We evaluate these methods on our data set using chest CT images. Table 8 shows the results of performance comparison with self-knowledge distillations.

From Table 8, it is apparent that our approach is better than others. Specifically, ours is 0.036, 0.058, and 0.122 higher than Ours (CS-KD), Ours (BYOT), and Ours (DDGSD), in

TABLE 6 Classification results of each model on COVID v2.0 data set

Methods		Accuracy	Precision	Sensitivity	Specificity	F1-score	AUC
Ours	Unsupervised	**0.999**	**0.893**	**1.000**	**0.999**	**0.909**	**0.987**
Ours w/o S-KD	Unsupervised	0.997	0.800	0.800	**0.999**	0.800	0.895
COVID-CAPS[101]	Supervised	0.957	0.823	0.900	0.958	0.860	0.970
COVID-ResNet[102]	Supervised	0.954	0.708	0.973	0.935	0.820	0.923
AI4COVID-19[103]	Supervised	0.889	0.875	0.833	0.933	0.903	0.887
DenseNet-121[91]	Supervised	0.925	0.882	0.900	0.921	0.906	0.898
DenseNet-161[92]	Supervised	0.805	0.828	0.860	0.841	0.843	0.832
ResNet-34[93]	Supervised	0.835	0.840	0.860	0.849	0.851	0.825
VGG-19[94]	Supervised	0.818	0.850	0.830	0.813	0.845	0.808
ResNet-18[95]	Supervised	0.848	0.860	0.820	0.783	0.839	0.786
EfficientNet-B0[96]	Supervised	0.799	0.806	0.825	0.772	0.788	0.848
EfficientNet-B1[96]	Supervised	0.696	0.686	0.690	0.651	0.698	0.772
EfficientNet-B2[96]	Supervised	0.759	0.719	0.712	0.680	0.725	0.794
EfficientNet-B3[96]	Supervised	0.735	0.757	0.762	0.709	0.706	0.793
EfficientNet-B4[96]	Supervised	0.759	0.770	0.764	0.778	0.750	0.883
EfficientNet-B5[96]	Supervised	0.813	0.769	0.763	0.793	0.816	0.845
Inception-v3[97]	Supervised	0.897	0.843	0.854	0.878	0.894	0.941
Inception-ResNet-v2[98]	Supervised	0.864	0.882	0.821	0.816	0.894	0.936
MobileNet-v2[99]	Supervised	0.861	0.890	0.949	0.757	0.863	0.949
DenseNet-201[100]	Supervised	0.749	0.790	0.760	0.704	0.783	0.724
VGG-16[94]	Supervised	0.717	0.720	0.750	0.699	0.734	0.697

terms of accuracy, respectively. In terms of precision, sensitivity, specificity, F1-Score, AUC, there are similar scenarios as the above. *Thus, our approach is more effective than other state-of-the-art self-knowledge distillation methods for COVID-19 diagnosis from pneumonia cases.*

5.5 | Compared to different loss functions in feature extraction

To better verify the effective performance of our loss function in feature extraction, we compare our model with other state-of-the-art loss functions, including Proxy-NCA,[106] SoftTriple,[80] Triplet,[79] N-pair,[107] and Lifted Structure.[108] "Ours (Proxy-NCA)" means a variant of Ours, which only using Proxy-NCA and not using our feature extraction loss funtion (i.e., $\mathcal{L}_{\text{Cos-Triplet}}$). "Ours (SoftTriple)" means a variant of Ours, which only using SoftTriple and not using our feature extraction loss funtion. "Ours (Triplet)" means a variant of Ours, which only using Triplet and not using our feature extraction loss funtion. "Ours(N-pair)" means a variant of Ours, which only using N-pair and not using our feature extraction loss funtion. "Ours (Lifted Structure)" means a variant of Ours, which only using Lifted Structure and not using our

TABLE 7 Classification results of each model on COVID CT data set

Methods		Accuracy	Precision	Sensitivity	Specificity	F1-score	AUC
Ours	Unsupervised	**0.980**	**0.990**	**0.969**	**0.990**	**0.979**	**0.989**
Ours w/o S-KD	Unsupervised	0.951	0.949	0.949	0.952	0.949	0.967
DenseNet-121[91]	Supervised	0.935	0.920	0.939	0.935	0.936	0.937
DenseNet-161[92]	Supervised	0.810	0.887	0.867	0.844	0.877	0.836
ResNet-34[93]	Supervised	0.842	0.847	0.865	0.851	0.856	0.827
VGG-19[94]	Supervised	0.825	0.853	0.830	0.813	0.842	0.812
ResNet-18[95]	Supervised	0.852	0.865	0.826	0.791	0.845	0.794
EfficientNet-B0[96]	Supervised	0.820	0.847	0.829	0.842	0.815	0.907
EfficientNet-B1[96]	Supervised	0.710	0.727	0.723	0.690	0.712	0.809
EfficientNet-B2[96]	Supervised	0.770	0.768	0.760	0.748	0.768	0.859
EfficientNet-B3[96]	Supervised	0.760	0.769	0.762	0.721	0.763	0.851
EfficientNet-B4[96]	Supervised	0.790	0.791	0.784	0.778	0.788	0.877
EfficientNet-B5[96]	Supervised	0.820	0.817	0.801	0.805	0.817	0.886
Inception-v3[97]	Supervised	0.913	0.885	0.903	0.924	0.945	0.970
Inception-ResNet-v2[98]	Supervised	0.864	0.902	0.882	0.843	0.923	0.950
MobileNet-v2[99]	Supervised	0.873	0.922	0.958	0.776	0.915	0.950
DenseNet-201[100]	Supervised	0.749	0.792	0.765	0.710	0.778	0.725
VGG-16[94]	Supervised	0.720	0.720	0.751	0.705	0.735	0.702

feature extraction loss funtion. We evaluate these methods on our data set using chest CT images. Table 9 shows the results of performance comparison with loss functions.

From Table 9, it is apparent that our approach is better than others. Specifically, ours is 0.139, 0.109, 0.080, 0.081, and 0.135 higher than Ours (Proxy-NCA), Ours (SoftTriple), Ours (Triplet), Ours (N-pair), and Ours (Lifted Structure), in term of accuracy, respectively. In terms of precision, sensitivity, specificity, F1-Score, AUC, there are similar scenarios as the above. *Thus, our approach is more effective than other state-of-the-art losses for COVID-19 diagnosis from pneumonia cases.*

TABLE 8 Comparison results of each self-knowledge distillation on our data set using chest CT images

	Accuracy	Precision	Sensitivity	Specificity	F1-Score	AUC
Ours	**0.986**	**0.990**	**0.987**	**0.986**	**0.988**	**0.991**
Ours (CS-KD[35])	0.950	0.925	0.874	0.955	0.899	0.929
Ours (BYOT[104])	0.928	0.949	0.863	0.967	0.904	0.872
Ours (DDGSD[105])	0.864	0.952	0.854	0.980	0.900	0.947

TABLE 9 Comparison results of each loss function in feature extraction on our data set using chest CT images

Methods	Accuracy	Precision	Sensitivity	Specificity	F1-Score	AUC
Ours	**0.986**	**0.990**	**0.987**	**0.986**	**0.988**	**0.991**
Ours (Proxy-NCA[106])	0.847	0.970	0.867	0.887	0.916	0.920
Ours (SoftTriple[80])	0.877	0.856	0.940	0.943	0.896	0.838
Ours (Triplet[79])	0.906	0.878	0.917	0.857	0.897	0.887
Ours (N-pair[107])	0.905	0.958	0.868	0.857	0.911	0.805
Ours (Lifted Structure[108])	0.851	0.889	0.837	0.747	0.862	0.802

5.6 | Compared to different networks in feature extraction

To better verify the effective performance of our network in feature extraction, we compare our model with other state-of-the-art network, including EfficientNet-B4,[96] Inception-v3,[97] Inception-ResNet-v2,[98] MobileNet-v2,[99] ResNet-50,[109] and VGG-16.[110] "Ours (EfficientNet-B4)" means a variant of Ours, which using EfficientNet-B4 and not using our network. "Ours (Inception-v3)" means a variant of Ours, which using Inception-v3 and not using our network. "Ours (Inception-ResNet-v2)" means a variant of Ours, which using Inception-ResNet-v2 and not using our network. "Ours (MobileNet-v2)" means a variant of Ours, which using MobileNet-v2 and not using our network. "Ours (ResNet-50)" means a variant of Ours, which using ResNet-50 and not using our network. "Ours (VGG-16)" means a variant of Ours, which using VGG-16 and not using our network. We evaluate these methods on our data set using chest CT images. Table 10 shows the results of performance comparison with self-knowledge distillations.

From Table 10, it is apparent that our approach is better than others. Specifically, ours is 0.231, 0.095, 0.127, 0.131, 0.143, and 0.076 higher than Ours (EfficientNet-B4), Ours (Inception-v3), Ours (Inception-ResNet-v2), Ours (MobileNet-v2), Ours (ResNet-50), Ours (VGG-16), in term of accuracy, respectively. In term of precision, sensitivity, specificity, F1-Score, AUC, there are similar scenarios as the above. *Thus, our approach is more effective than other state-of-the-art networks for COVID-19 diagnosis from pneumonia cases.*

TABLE 10 Comparison results of each network in feature extraction on our data set using chest CT images

Methods	Accuracy	Precision	Sensitivity	Specificity	F1-Score	AUC
Ours	**0.986**	**0.990**	**0.987**	**0.986**	**0.988**	**0.991**
Ours (EfficientNet-B4[96])	0.755	0.765	0.763	0.775	0.748	0.903
Ours (Inception-v3[97])	0.891	0.838	0.850	0.876	0.891	0.924
Ours (Inception-ResNet-v2[98])	0.859	0.877	0.818	0.814	0.892	0.911
Ours (MobileNet-v2[99])	0.855	0.886	0.946	0.755	0.860	0.912
Ours (ResNet-50[109])	0.843	0.949	0.923	0.848	0.936	0.949
Ours (VGG-16[110])	0.911	0.939	0.939	0.882	0.939	0.949

5.7 | Compared to different relation models

To better verify the effective performance of our relation model, we compare our model with other state-of-the-art relation models, including the relation models from Zheng-MICCAI,[111] and Zheng-ICME.[112] "Ours (Zheng-MICCAI)" means a variant of Ours, which using relation models from Zheng-MICCAI[111] and not using our relation model. "Ours (Zheng-ICME)" means a variant of Ours, which using relation models from Zheng-ICME[112] and not using our relation model. We evaluate these methods on our data set using chest CT images. Table 11 shows the results of performance comparison with self-knowledge distillations.

From Table 11, it is apparent that our approach is better than others. Specifically, ours is 0.111 and 0.077 higher than Ours (Zheng-MICCAI) and Ours (Zheng-ICME), in terms of accuracy, respectively. In terms of precision, sensitivity, specificity, F1-Score, AUC, there are similar scenarios as the above. *Thus, our approach is more effective than other state-of-the-art networks for COVID-19 diagnosis from pneumonia cases.*

5.8 | Discussion about different meta-learning

In this section, we compare with state-of-the-art meta-learning approaches to verify the effectiveness of ours. We evaluate these methods on our data set using chest CT images. Table 12 shows the results of performance comparison with different meta-learnings.

We can divide meta-learning methods into three categories[22]:

(1) Metric learning methods (i.e., MatchingNets,[121] ProtoNets,[122] RelationNets,[123] Graph neural network (GraphNN),[124] Ridge regression,[125] TransductiveProp[126]), FEAT[127] learn a similarity space in which learning is particularly efficient for few-shot examples.
(2) Memory network methods (i.e., Meta Networks,[113] TADAM[114]) learn to store "experience" when learning seen tasks and then generalize it to unseen tasks.
(3) Gradient descent based meta-learning methods (i.e., CACTUs,[47] UMTRA,[48] MAML,[115] Meta-LSTM,[116] MetaGAN,[117] LEO,[118] LGM-Net,[119] CTM,[120] and WarpGrad[90]) intend for adjusting the optimization algorithm so that the model can converge within a small number of optimization steps (with a few examples).

From above and Table 12, we can get the following three points:

First, metric-based methods propose that samples of the same class are close to each other, and samples of the different classes are far away from each other by simulating the metric distribution among samples. Generally speaking, the neural network is used to construct the embedding space (feature space) of samples, and some measure is used to calculate the similarity between samples. There are inaccurate (or noisy) labels of these images that can

TABLE 11 Comparison results of each relation model on our data set using chest CT images

Methods	Accuracy	Precision	Sensitivity	Specificity	F1-Score	AUC
Ours	**0.986**	**0.990**	**0.987**	**0.986**	**0.988**	**0.991**
Ours (Zheng-MICCAI[111])	0.875	0.918	0.885	0.856	0.901	0.899
Ours (Zheng-ICME[112])	0.909	0.919	0.893	0.897	0.906	0.838

TABLE 12 Comparison results of different meta-learning on our data set using chest CT images

Meta-learning method			Accuracy	Precision	Sensitivity	Specificity	F1-Score	AUC
Memory network	Supervised	Meta Networks[113]	0.796	0.835	0.806	0.809	0.820	0.769
	Supervised	TADAM[114]	0.801	0.842	0.825	0.813	0.834	0.789
Gradient descent	*Unsupervised*	CACTUs[47]	0.788	0.808	0.791	0.806	0.800	0.749
	Unsupervised	UMTRA[48]	0.789	0.822	0.803	0.807	0.812	0.758
	Supervised	MAML[115]	0.802	0.843	0.829	0.815	0.836	0.797
	Supervised	Meta-LSTM[116]	0.825	0.849	0.830	0.820	0.840	0.821
	Supervised	MetaGAN[117]	0.834	0.895	0.833	0.830	0.863	0.837
	Supervised	LEO[118]	0.834	0.896	0.836	0.835	0.865	0.838
	Supervised	LGM-Net[119]	0.870	0.903	0.863	0.862	0.883	0.871
	Supervised	CTM[120]	0.878	0.913	0.865	0.866	0.888	0.881
	Supervised	WarpGrad[90]	0.905	0.922	0.881	0.935	0.901	0.915
Metric learning	Supervised	MatchingNets[121]	0.803	0.843	0.829	0.818	0.836	0.818
	Supervised	ProtoNets[122]	0.826	0.864	0.832	0.822	0.848	0.829
	Supervised	RelationNets[123]	0.841	0.898	0.850	0.836	0.873	0.839
	Supervised	Graph neural network[124]	0.843	0.899	0.857	0.838	0.877	0.847
	Supervised	Ridge regression[125]	0.844	0.903	0.862	0.839	0.882	0.861
	Supervised	TransductiveProp[126]	0.872	0.913	0.864	0.865	0.888	0.874
	Supervised	FEAT[127]	0.905	0.929	0.884	0.940	0.906	0.925
	Unsupervised	Ours	**0.986**	**0.990**	**0.987**	**0.986**	**0.988**	**0.991**

influence the recognition of three kinds of data sets. The existing metric-based methods do not consider these kinds of features in this scenario. Therefore, in the feature space of chest CT images, most of the features are unserviceable, and similarities among these images cannot be performed effectively.

Second, gradient descent-based methods mostly directly optimize an initial feature representation. Based on this feature representation, the model can be efficiently adjusted using gradient updating based on a few images. However, in the task of COVID-19 diagnosis from pneumonia cases in this paper, even though the gradient descent based method may initially learn the resultful features of some images, most chest CT images with a large number of inaccurate (or noisy) labels can cause this kind of method to fail to adjust gradient in subsequent training. As a result, the accuracy and the generalization ability of this kind of method decrease.

Third, the memory network methods have an architecture that enhances memory capacity, which provides the ability to encode and retrieve new information quickly. In other words, this kind of method focuses on what is in memory capacity (memory network). Due to a large number of inaccurate (or noisy) labels in chest CT images, most of the memory capacity is useless features. As a result, the accuracy and generalization ability of this kind of method is reduced in the COVID-19 diagnosis from pneumonia cases task.

All in all, from Table 12, ours is better than others. *From the above discussion, it is clear that our approach is more effective than state-of-the-art meta-learning approaches.*

6 | CONCLUSIONS

COVID-19 pandemic is a significant contributor to the overall global burden of diseases now. The state-of-the-art COVID-19 diagnosis approaches employ deep-learning-based networks to obtain a robust model. However, these models, which only focus on distinguishing COVID-19 or not, suffer from the problem of robustness resulting from a few public data about COVID-19 pneumonia and ignore the importance of employing a model's knowledge distillation to improve the task of performance.

In this paper, we propose a new model for COVID-19 diagnosis from pneumonia cases, which can use the style of meta-learning without labels and fully distill the knowledge of the model itself to improve the performance. By being its own teacher, our approach not only obtains the accurate feature embeddings of medical information but also directly refines the medical knowledge from itself. With the learned feature embeddings and meta-learning-based networks, our approach can learn to discriminate the images and learn to judge whether images belongs to COVID-19 cases. Experimental results on three real-world data sets validate the effectiveness and robustness of the proposed approach. Besides, a new data set is constructed by utilizing text mining from medical reports. Compared with previous data sets, our data set not only contains more chest images from COVID-19 patients but also collect chest images from diagnosed patients with other types of pneumonia that are similar to the symptoms of COVID-19 and the fine-grained labels of all. We hope this study will inspire others to build artificial intelligence-based tools to accelerate the anti-epidemic of COVID-19.

ACKNOWLEDGMENTS

We would like to thank the General Hospital of the People's Liberation Army and Wuhan Pulmonary Hospital for medical data and helpful advice on this study. This study is supported

in part by the National Key R&D Program of China (2020YFB1600400), in part by the National Natural Science Foundation of China (61806198, 61533019, and U1811463), in part by the Key Research and Development Program of Guangzhou (202007050002), and in part by the National Key Research and Development Program of China (No. 2018AAA0101502).

CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

AUTHOR CONTRIBUTIONS

Wenbo Zheng: Conceptualization, methodology, software, validation, investigation, writing—original draft, writing—review and editing; **Lan Yan**: Methodology, investigation, writing—original draft; **Chao Gou**: Conceptualization, resources, writing—review and editing, project administration, funding acquisition; **Zhi-Cheng Zhang**: Validation, software, resources; **Jun J. Zhang**: Supervision, software, validation, investigation; **Ming Hu**: Validation, software, resource; **Fei-Yue Wang**: Supervision and funding acquisition.

ENDNOTES

* https://github.com/ShahinSHH/COVID-CAPS

† https://github.com/lindawangg/COVID-Net

‡ https://github.com/adeaeede/ai4covid

ORCID

Wenbo Zheng https://orcid.org/0000-0001-9732-3217
Lan Yan https://orcid.org/0000-0001-6452-9649
Chao Gou https://orcid.org/0000-0002-4128-886X
Zhi-Cheng Zhang https://orcid.org/0000-0001-9768-9470
Jun J. Zhang https://orcid.org/0000-0003-3792-9510
Ming Hu https://orcid.org/0000-0003-0295-226X
Fei-Yue Wang https://orcid.org/0000-0001-9185-3989

REFERENCES

1. Haar v.dJ, Hoes LR, Coles CE, et al. Caring for patients with cancer in the COVID-19 era. *Nat Med*. 2020;26(5):665-671. https://doi.org/10.1038/s41591-020-0874-8
2. Jia JS, Lu X, Yuan Y, Xu G, Jia J, Christakis NA. Population flow drives spatio-temporal distribution of COVID-19 in China. *Nature*. 2020;582:389-394. https://doi.org/10.1038/s41586-020-2284-y
3. Reynolds HR, Adhikari S, Pulgarin C. Renin-angiotensin-aldosterone system inhibitors and risk of Covid-19. *New England J Med*. 2020;382:2441-2448. https://doi.org/10.1056/NEJMoa2008975
4. Zhang X, Tan Y, Ling Y, et al. Viral and host factors related to the clinical outcome of COVID-19. *Nature*. 2020. https://doi.org/10.1038/s41586-020-2355-0
5. Bojkova D, Klann K, Koch B, et al. Proteomics of SARS-CoV-2-infected host cells reveals therapy targets. *Nature*. 2020;583:469-472. https://doi.org/10.1038/s41586-020-2332-7
6. Mehra MR, Desai SS, Kuy S, Henry TD, Patel AN. Cardiovascular disease, drug therapy, and mortality in Covid-19. *New England J Med*. 2020;382:e102. https://doi.org/10.1056/NEJMoa2007621
7. Ikizler TA, Kliger AS. Minimizing the risk of COVID-19 among patients on dialysis. *Nat Rev Nephrol*. 2020;16:311-313. https://doi.org/10.1038/s41581-020-0280-y
8. Lai S, Ruktanonchai NW, Zhou L, et al. Effect of non-pharmaceutical interventions to contain COVID-19 in China. *Nature*. 2020. https://doi.org/10.1038/s41586-020-2293-x

9. Geleris J, Sun Y, Platt J, et al. Observational study of hydroxychloroquine in hospitalized patients with Covid-19. *New England J Med*. 2020;382:2411-2418. https://doi.org/10.1056/NEJMoa2012410
10. Shi F, Wang J, Shi J, et al. Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for COVID-19. *IEEE Rev Biomed Eng*. 2021;14:4-15. https://doi.org/10.1109/RBME.2020.2987975
11. Yan L, Zhang HT, Goncalves J, et al. An interpretable mortality prediction model for COVID-19 patients. *Nat Mach Intell*. 2020;2(5):283-288. https://doi.org/10.1038/s42256-020-0180-7
12. Mei X, Lee HC, Diao Ky, et al. Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. *Nat Med*. 2020;26:1224-1228. https://doi.org/10.1038/s41591-020-0931-3
13. Wölfel R, Corman VM, Guggemos W, et al. Virological assessment of hospitalized patients with COVID-2019. *Nature*. 2020;581:465-469. https://doi.org/10.1038/s41586-020-2196-x
14. Liang W, Yao J, Chen A, et al. Early triage of critically ill COVID-19 patients using deep learning. *Nat Commun*. 2020;11(1):3543. https://doi.org/10.1038/s41467-020-17280-8
15. Dai L, Zheng T, Xu K, et al. A universal design of betacoronavirus vaccines against COVID-19, MERS, and SARS. *Cell*. 2020;182:722-733. https://doi.org/10.1016/j.cell.2020.06.035
16. Liu WJ, Lan J, Liu K, et al. Protective T cell responses featured by concordant recognition of middle east respiratory syndrome coronavirus-Derived CD8. T Cell Epitopes and Host MHC. *J Immunol*. 2017;198(2):873-882. https://doi.org/10.4049/jimmunol.1601542
17. Zhou M, Xu D, Li X, et al. Screening and identification of severe acute respiratory syndrome-associated coronavirus-specific CTL epitopes. *J Immunol*. 2006;177(4):2138-2145. https://doi.org/10.4049/jimmunol.177.4.2138
18. Banerjee A, Kulcsar K, Misra V, Frieman M, Mossman K. Bats and coronaviruses. *Viruses*. 2019;11(1):41. https://doi.org/10.3390/v11010041
19. Wong AC, Li X, Lau SK, Woo PC. Global epidemiology of bat coronaviruses. *Viruses*. 2019;11(2):174. https://doi.org/10.3390/v11020174
20. Weitz JS, Beckett SJ, Coenen AR, et al. Modeling shield immunity to reduce COVID-19 epidemic spread. *Nat Med*. 2020;26:849-854. https://doi.org/10.1038/s41591-020-0895-3
21. Long QX, Liu BZ, Deng HJ, et al. Antibody responses to SARS-CoV-2 in patients with COVID-19. *Nat Med*. 2020;26:845-848. https://doi.org/10.1038/s41591-020-0897-1
22. Sun Q, Liu Y, Chua T, Schiele B. Meta-transfer learning for few-shot learning. 2019:403-412.
23. Passalis N, Iosifidis A, Gabbouj M, Tefas A. Hypersphere-based weight imprinting for few-shot learning on embedded devices. *IEEE Trans Neural Networks Learn Syst*. 2021;32(2):925-930. https://doi.org/10.1109/TNNLS.2020.2979745
24. Jung HG, Lee SW. Few-shot learning with geometric constraints. *IEEE Trans Neural Network Learn Syst*. 2020;31(11):4660-4672. https://doi.org/10.1109/TNNLS.2019.2957187
25. Guan Wj, Ni Zy, Hu Y, et al. Clinical characteristics of coronavirus disease 2019 in China. *New England J Med*. 2020;382(18):1708-1720. https://doi.org/10.1056/NEJMoa2002032
26. Pang J, Tao F, Huang Q, Tian Q, Yin B. Two birds with one stone: A coupled poisson deconvolution for detecting and describing topics from multimodal web data. *IEEE Trans Neural Network Learn Syst*. 2019;30(8):2397-2409.
27. Niu L, Xu X, Chen L, Duan L, Xu D. Action and event recognition in videos by learning from heterogeneous web sources. *IEEE Trans Neural Network Learn Syst*. 2017;28(6):1290-1304. http://doi.org/10.1109/tnnls.2016.2518700
28. Zhang J, Xie Y, Li Y, Shen C, Xia Y. COVID-19 screening on chest X-ray images using deep learning based anomaly detection. arXiv preprint arXiv:2003.12338; 2020.
29. Wang L, Wong A. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images. *arXiv preprint arXiv*:2003.09871; 2020.
30. Jin C, Chen W, Cao Y, et al. Development and evaluation of an AI system for COVID-19 diagnosis. *medRxiv*. 2020.
31. Wang Y, Yao Q, Kwok JT, Ni LM. Generalizing from a few examples: A survey on few-shot learning. *ACM Comput Surv*. 2020;53(3). https://doi.org/10.1145/3386252
32. Hospedales T, Antoniou A, Micaelli P, Storkey A. Meta-learning in neural networks: A survey. arXiv preprint arXiv:2004.05439; 2020.

33. Gu X, Angelov PP, Soares EA. A self-adaptive synthetic over-sampling technique for imbalanced classification. *Int J Intell Syst*. 2020;35(6):923-943. https://doi.org/10.1002/int.22230
34. Tian L, Zheng D, Zhu C. Image classification based on the combination of text features and visual features. *Int J Intell Syst*. 2013;28(3):242-256. https://doi.org/10.1002/int.21567
35. Yun S, Park J, Lee K, Shin J. Regularizing class-wise predictions via self-knowledge distillation. 2020.
36. Asano YM, Patrick M, Rupprecht C, Vedaldi A. Labelling unlabelled videos from scratch with multimodal self-supervision. 2020.
37. Lee DH, et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. 2013.
38. Yalniz IZ, Jégou H, Chen K, Paluri M, Mahajan D, Billion-scale semi-supervised learning for image classification. *CoRR*; 2019, abs/1905.00546.
39. Xie Q, Luong MT, Hovy E, Le QV. Self-training with noisy student improves imagenet classification. 2020.
40. Zoph B, Ghiasi G, Lin TY, et al. Rethinking pre-training and self-training. arXiv preprint arXiv:2006.06882; 2020.
41. He J, Gu J, Shen J, Ranzato M. Revisiting self-training for neural sequence generation. 2020.
42. Kahn J, Lee A, Hannun A. Self-training for end-to-end speech recognition. 2020:7084-7088.
43. Park DS, Zhang Y, Jia Y, et al. Improved noisy student training for automatic speech recognition. 2020: 2817-2821.
44. Ge Y, Chen D, Li H. Mutual mean-teaching: pseudo label refinery for unsupervised domain adaptation on person re-identification. 2020.
45. Zhang Z, Zhang H, Arik SO, Lee H, Pfister T. Distilling effective supervision from severe label noise. 2020.
46. Ji Z, Zou X, Huang T, Wu S. Unsupervised few-shot learning via self-supervised training. arXiv preprint arXiv:1912.12178; 2019.
47. Hsu K, Levine S, Finn C. Unsupervised learning via meta-learning. 2019.
48. Khodadadeh S, Boloni L, Shah M. Unsupervised meta-learning for few-shot image classification. Curran Associates, Inc.; 2019:10132-10142.
49. Antoniou A, Storkey A. Assume, augment and learn: Unsupervised few-shot meta-learning via random labels and data augmentation. arXiv preprint arXiv:1902.09884; 2019.
50. TiexinQin YS, Yang G. Unsupervised few-shot learning via distribution shift-based augmentation. 2020.
51. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-Ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. 2017: 3462-3471.
52. Zhang S, Jing L, Tao J, Qi L, Jianping L. Application of imaging examination in diagnosis of viral pneumonia. *Int J Respir*. 2020. https://doi.org/10.3760/cma.j.cn131368-20200409-00266
53. Kanne JP. Chest CT Findings in 2019 Novel Coronavirus (2019-nCoV) Infections from Wuhan, China: Key Points for the Radiologist. *Radiology*. 2020;295(1):16-17. PMID: 3201766210.1148/radiol.2020200241
54. Chung M, Bernheim A, Mei X, et al. CT imaging features of 2019 novel coronavirus (2019-nCoV). *Radiology*. 2020;295(1):202-207.
55. Lei J, Li J, Li X, Qi X. CT imaging of the 2019 novel coronavirus (2019-nCoV) Pneumonia. *Radiology*. 2020; 295(1):18.
56. Ma Y, Zhang S, Zhao L, et al. Inhalation lung injury induced by smoke bombs in children: CT manifestations, dynamic evolution features and quantitative analysis. *J Thoracic Disease*. 2018;10:10.
57. Tan D, Fu Y, Xu J, et al. Severe adenovirus community-acquired pneumonia in immunocompetent adults: chest radiographic and CT findings. *J Thoracic Disease*. 2016;8:5.
58. Das KM, Lee EY, Jawder SEA, et al. Acute middle east respiratory syndrome coronavirus: Temporal lung changes observed on the chest radiographs of 55 patients. *Am J Roentgenol*. 2015;205(3):W267-S274. https://doi.org/10.2214/AJR.15.14445
59. AdÃńno IFA, Tibana TK, Santos RAFATA, et al. Initial chest X-ray findings in pediatric patients diagnosed with H1N1 virus infection. *Radiologia Brasileira*. 2019;52:78-84.
60. Olson G, Davis AM. Diagnosis and treatment of adults with community-acquired Pneumonia. *JAMA*. 2020;323(9):885-886. https://doi.org/10.1001/jama.2019.21118
61. Mo X, Jian W, Su Z, et al. Abnormal pulmonary function in COVID-19 patients at time of hospital discharge. *Eur Respir J*. 2020;55(6). https://doi.org/10.1183/13993003.01217-2020

62. Chen R, Sang L, Jiang M, et al. Longitudinal hematologic and immunologic variations associated with the progression of COVID-19 patients in China. *J Allergy Clin Immunol*. 2020;146(1):89-100. https://doi.org/10.1016/j.jaci.2020.05.003
63. Sun J, Zhuang Z, Zheng J, et al. Generation of a broadly useful model for COVID-19 pathogenesis, vaccination, and treatment. *Cell*. 2020;182(3):734-743. https://doi.org/10.1016/j.cell.2020.06.010
64. Zheng R, Zhou J, Song B, et al. COVID-19-associated coagulopathy: thromboembolism prophylaxis and poor prognosis in ICU. *Exp Hematol Oncol*. 2021; 10(1): 6. https://doi.org/10.1186/s40164-021-00202-9
65. Liang W, Liang H, Ou L, et al. Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with COVID-19. *JAMA Internal Med*. 2020;180(8): 1081-1089. https://doi.org/10.1001/jamainternmed.2020.2033
66. Zhang W, Zhou T, Lu Q, et al. Dynamic fusion-based federated learning for COVID-19 detection. *IEEE Internet Things J*. 2021: 1. https://doi.org/10.1109/JIOT.2021.3056185
67. Ohata EF, Bezerra GM, ChagasdJVS, et al. Automatic detection of COVID-19 infection using chest X-ray images through transfer learning. *IEEE/CAA J Automatica Sinica*. 2021;8(1):239-248. https://doi.org/10.1109/JAS.2020.1003393
68. Chen C, Zhou K, Zha M, et al. An effective deep neural network for lung lesions segmentation from COVID-19 CT images. *IEEE Trans Industrial Inform*. 2021:1. https://doi.org/10.1109/TII.2021.3059023
69. Shamsi A, Asgharnezhad H, Jokandan SS, et al. An uncertainty-aware transfer learning-based framework for COVID-19 diagnosis. *IEEE Trans Neural Network Learn Syst*. 2021:1-10. https://doi.org/10.1109/TNNLS.2021.3054306
70. Li J, Wang Y, Wang S, et al. Multiscale attention guided network for COVID-19 diagnosis using chest X-ray images. *IEEE J Biomed Health Inform*. 2021:1. https://doi.org/10.1109/JBHI.2021.3058293
71. Tang S, Wang C, Nie J, et al. EDL-COVID: Ensemble deep learning for COVID-19 cases detection from chest X-ray images. *IEEE Trans Industrial Inform*. 2021:1. https://doi.org/10.1109/TII.2021.3057683
72. Castiglione A, Vijayakumar P, Nappi M, Sadiq S, Umer M. COVID-19: automatic detection of the novel coronavirus disease from CT images using an optimized convolutional neural network. *IEEE Trans Industrial Inform*. 2021:1. https://doi.org/10.1109/TII.2021.3057524
73. Paluru N, Dayal A, Jenssen HB, et al. Anam-Net: anamorphic depth embedding-based lightweight CNN for segmentation of anomalies in COVID-19 chest CT images. *IEEE Trans Neural Network Learn Syst*. 2021:1-15. https://doi.org/10.1109/TNNLS.2021.3054746
74. Wang YX, Hebert M. Learning to learn: Model regression networks for easy small sample learning. Springer; 2016:616-634.
75. Wang YX, Ramanan D, Hebert M. Learning to model the tail. 2017:7029-7039.
76. Wang YX, Ramanan D, Hebert M. Meta-learning to detect rare objects. 2019:9925-9934.
77. Cohen JP, Morrison P, Dao L. COVID-19 image data collection. arXiv preprint arXiv:2003.11597; 2020.
78. Zhao J, Zhang Y, He X, Xie P. COVID-CT-dataset: a CT scan dataset about COVID-19. arXiv preprint arXiv:2003.13865; 2020.
79. Schroff F, Kalenichenko D, Philbin J. FaceNet: a unified embedding for face recognition and clustering. 2015.
80. Qian Q, Shang L, Sun B, Hu J, Li H, Jin R. SoftTriple loss: deep metric learning without triplet sampling. 2019.
81. Liu W, Wen Y, Yu Z, Yang M. Large-margin softmax loss for convolutional neural networks. In: 48 of *Proceedings of Machine Learning Research*. PMLR; New York, NY; 2016:507-516.
82. Isensee F, Petersen J, Kohl SA, Jäger PF, Maier-Hein KH. nnu-net: breaking the spell on successful medical image segmentation. arXiv preprint arXiv:1904.08128; 2019;1:1-8.
83. Faes L, Wagner SK, Fu DJ, et al. Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study. *Lancet Digital Health*. 2019;1(5): e232-e242. https://doi.org/10.1016/S2589-7500(19)30108-6
84. Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. 2015.
85. Passalis N, Tefas A. Unsupervised knowledge transfer using similarity embeddings. *IEEE Trans Neural Network Learn Syst*. 2019;30(3):946-950. https://doi.org/10.1109/TNNLS.2018.2851924
86. Chen H, Wang Y, Xu C, Xu C, Tao D. Learning student networks via feature embedding. *IEEE Trans Neural Network Learn Syst*. 2020:1-11.

87. Guo T, Xu C, He S, Shi B, Xu C, Tao D. Robust student network learning. *IEEE Transactions on Neural Networks and Learning Systems*. 2019: 1-14.
88. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. In; 2017.
89. Iandola F, Moskewicz M, Karayev S, Girshick R, Darrell T, Keutzer K. Densenet: Implementing efficient convnet descriptor pyramids. arXiv preprint arXiv:1404.1869; 2014.
90. Flennerhag S, Rusu AA, Pascanu R, Visin F, Yin H, Hadsell R. Meta-learning with warped gradient descent. 2020.
91. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. 2017: 4700-4708.
92. SDCT-AuxNet: DCT augmented stain deconvolutional CNN with auxiliary classifier for cancer diagnosis. Medical Image Analysis. 2020;61:101661.
93. Guo H, Kruger U, Wang G, Kalra MK, Yan P. Knowledge-based analysis for mortality prediction from CT images. *IEEE J Biomed Health Inform*. 2020;24(2):457-464.
94. Taylor AG, Mielke C, Mongan J. Automated detection of moderate and large pneumothorax on frontal chest X-rays using deep convolutional neural networks: a retrospective study. *PLOS Medicine*. 2018; 15.
95. An automatic method for lung segmentation and reconstruction in chest X-ray using deep neural networks. *Comput Methods Programs Biomed*. 2019;177:285-296.
96. Tan M, Le Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In: Chaudhuri K, Salakhutdinov R., eds. *Proceedings of the 36th International Conference on 97 Proceedings of Research*. PMLR; 2019:6105-6114.
97. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. 2016:2818-2826.
98. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In: AAAI'17. AAAI Press; 2017:4278-4284.
99. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: Inverted residuals and linear bottlenecks. 2018.
100. Nóbrega dRVM, Rebouças Filho PP, Rodrigues MB, Silva dSPP, Dourado Júnior CMJM, Albuquerque dVHC. Lung nodule malignancy classification in chest computed tomography images using transfer learning and convolutional neural networks. *Neural Comput Appl*. 2018.
101. Afshar P, Heidarian S, Naderkhani F, Oikonomou A, Plataniotis KN, Mohammadi A. COVID-CAPS: a capsule network-based framework for identification of COVID-19 cases from X-ray images. arXiv preprint arXiv:2004.02696; 2020.
102. Farooq M, Hafeez A. COVID-ResNet: a deep learning framework for screening of COVID19 from radiographs. arXiv preprint arXiv:2003.14395; 2020.
103. Imran A, Posokhova I, Qureshi HN, et al. AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. arXiv preprint arXiv:2004.01275; 2020.
104. Zhang L, Song J, Gao A, Chen J, Bao C, Ma K. Be your own teacher: improve the performance of convolutional neural networks via self distillation. 2019.
105. Xu TB, Liu CL. Data-distortion guided self-distillation for deep neural networks. 2019:5565-5572.
106. Movshovitz-Attias Y, Toshev A, Leung TK, Ioffe S, Singh S. No fuss distance metric learning using proxies. 2017.
107. Sohn K. Improved deep metric learning with multi-class n-pair loss objective. In: Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R., eds. *Advances in Neural Information Processing Systems*. Curran Associates, Inc.; 2016:1857-1865.
108. OhSong H, Xiang Y, Jegelka S, Savarese S. Deep metric learning via lifted structured feature embedding. 2016.
109. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016.
110. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2015.
111. Zheng W, Gou C, Yan L. A relation hashing network embedded with prior features for skin lesion classification. In: Suk HI, Liu M, Yan P, Lian C., eds. *Machine Learning in Medical Imaging*. Cham: Springer International Publishing; 2019:115-123.
112. Zheng W, Yan L, Gou C, Zhang W, Wang F. A relation network embedded with prior features for few-shot caricature recognition. 2019:1510-1515.

113. Munkhdalai T, Yu H. Meta Networks. In: 70 of *Proceedings of Machine Learning Research*. PMLR; International Convention Centre, Sydney, Australia; 2017:2554-2563.
114. Oreshkin B, RodríguezLópez P, Lacoste A. TADAM: Task dependent adaptive metric for improved few-shot learning. 2018:721-731.
115. Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: 70 of *Proceedings of Machine Learning Research*. International Convention Centre, Sydney, Australia; 2017: 1126-1135.
116. Ravi S, Larochelle H Optimization as a model for few-shot learning. In: OpenReview.net; 2017.
117. Zhang R, Che T, Ghahramani Z, Bengio Y, Song Y. MetaGAN: an adversarial approach to few-shot learning. 2018:2365-2374.
118. Rusu AA, Rao D, Sygnowski J, et al. Meta-learning with latent embedding optimization. 2019.
119. Li H, Dong W, Mei X, Ma C, Huang F, Hu BG. LGM-Net: Learning to generate matching networks for few-shot learning. In: 97 of *Proceedings of Machine Learning Research*. PMLR; Long Beach, CA; 2019: 3825-3834.
120. Li H, Eigen D, Dodge S, Zeiler M, Wang X. Finding task-relevant features for few-shot learning by category traversal. 2019.
121. Vinyals O, Blundell C, Lillicrap T, kavukcuoglu k, Wierstra D. Matching networks for one shot learning. 2016:3630-3638.
122. Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. In: Guyon I, Luxburg UV, Bengio S, eds. *Advances in Neural Information Processing Systems*. Vol 30, 2017:4077-4087.
123. Sung F, Yang Y, Zhang L, Xiang T, Torr PHS, Hospedales TM. Learning to compare: Relation network for few-shot learning. 2018:1199-1208.
124. Satorras VG, Estrach JB. Few-shot learning with graph neural networks. 2018.
125. Bertinetto L, Henriques JF, Torr P, Vedaldi A. Meta-learning with differentiable closed-form solvers. 2019.
126. Liu Y, Lee J, Park M, et al. Learning to propagate labels: transductive propagation network for few-shot learning. 2019.
127. Ye HJ, Hu H, Zhan DC, Sha F. Few-shot learning via embedding adaptation with set-to-set functions. 2020.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Zheng W, Yan L, Gou C, et al. , et al. Learning to learn by yourself: Unsupervised meta-learning with self-knowledge distillation for COVID-19 diagnosis from pneumonia cases. *Int J Intell Syst*. 2021;36:4033–4064. https://doi.org/10.1002/int.22449